Vehicle Re-Identification using Shifted Window Transformer

By

Faisal Imran

FALL-2018-MSCS 00000275391 SEECS

Supervisor

Dr. Muhammad Shahzad

Department of Computing

School of Electrical Engineering & Computer Science (SEECS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

(July 2022)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Vehicle Re-Identification Using Shifted Window Transformer" written by FAISAL IMRAN, (Registration No 00000275391), 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: 23-Jul-2022

HoD/Associate Dean: _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Approval

It is certified that the contents and form of the thesis entitled "Vehicle Re-Identification Using Shifted Window Transformer" submitted by FAISAL IMRAN have been found satisfactory for the requirement of the degree

Advisor : Dr. Muhammad Shahzad

Signature: _____

Date: 23-Jul-2022

Committee Member 1:Dr. Muhammad Imran Malik

Signature: _____

Date: 23-Jul-2022

Committee Member 2:Dr. Arsalan Ahmad

Signature: _____

Date: 22-Jul-2022

Committee Member 3:Dr. Muhammad Moazam Fraz

Signature: _____

Date: 23-Jul-2022

Dedication

Dedicated to my mother and father for their never-ending support.

Certificate of Originality

I hereby declare that this submission titled "Vehicle Re-Identification Using Shifted Window Transformer" 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:FAISAL IMRAN

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 the course of 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 would not 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.

Faisal Imran

Contents

List of Abbreviations and Symbols

Abbreviations

Re-ID	Re-Identification
CNN	Convolutional Neural Network
ViT	Vision Transformer

List of Tables

List of Figures

Abstract

Vehicle re-identification has become a key area in computer vision which has important applications in areas like vehicle tracking, commodity tracking, security tracking, etc. But Vehicle re-identification inherits two challenges; there are large variations between the similar classes, similar classes have different looks in different viewpoints, and slight differences between different classes, two different vehicles may look the same as companies are making vehicles with similar appearances. Various CNN-based Vehicle Re-ID have been proposed but CNN faces two main challenges; only contain the information of neighborhood and loss of information due to sampling and convolutions. Transformer have recently been introduced in vision community and achieving superior results because they can solve fundamental problems of CNNs by keep long term relation and not losing any information. A novel approach based on Shifted Window Transformer was introduced that in this thesis that uses transformers model to tackle the vehicle re-id problem and uses side information module to enhance the discrimination ability of the model. To the best of our knowledge Shifted Window Transformer have first-time been used as a backbone in vehicle re-identification task and results were promising, the model was evaluated on benchmark dataset VeRi.

Chapter 1

Introduction

In 2017, [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 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.

Nonetheless, there exist many issues that can hinder the solution of these problems like inferior 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. Vehicle re-identification is such an application that can help in spotting a vehicle from non-overlapping cameras that be used in different scenarios like surveillance, intelligent transportation, and security. There exists a similar problem (Person re-identification) which in terms of accuracy of models is more mature but models for person re-id cannot be directly used for vehicle

re-id making vehicle re-id more challenging problem. As vehicle re-id inherits two major problems: 1) when a vehicle is captured using different viewpoints, its appearance is different in different viewpoints. 2) Two different vehicles with the The same model and same color will always have a similar appearance when they are captured from a similar viewpoint. Yi Zhou and Ling Shao [2] said that "The subtle inter-instance discrepancy between images of different vehicles and the large intra-instance difference between images of the same vehicle make the matching problem unsatisfactorily addressed by existing vision models". So, the idea is to train a model that can reduce the intra-instance difference between images of similar vehicles.

Model for vehicle and person re-identification contains three major parts including, feature learning, features extraction, distance matric learning, and calculation of distance based on matric learning. Feature extraction includes using a deep learning model to extract features from the images. Once these features are extracted, a deep learning model is trained to learn the representation from these features and make a distance matric which will future be used to address the vehicle re-id problem.

1.1 Background

Vehicle re-identification is a widespread problem that helps in solving problems like intelligent traffic management and security surveillance. These problems are the core problems in urban areas. Vehicle re-identification means the detection of the vehicle in multiple non-overlapping cameras. For example, there are 10 different cameras at a location and each camera is placed at a different view angle that is not overlapped with any other camera, vehicle re-id aims to detect a particular vehicle in all cameras at any point. Vehicle re-id is remarkably like the other field that is much more mature in terms of accuracy of the models and that is person re-id. Person re-id is the same problem but with complexity as in different viewpoints, the appearance of a person does not change drastically. For example, the texture and color of clothes are not going to change even in

large viewpoint variations, but they cannot be applied to a vehicle. In the case of the vehicle, the appearance of the vehicle in different viewpoints gets changed up to a substantial extent, making it harder for a model to differentiate between two vehicles. Another challenge in vehicle re-id is that if there are two vehicles with the same model and the same color, the difference between vehicles would be non-existent.

To tackle these problems of the same model vehicles, other methods like license plate and spatial-temporal information were used and the results were promising. However, in these two methods, the amount of information required to get these results was on the higher side. Especially for license plate detection, images must have a high resolution which is not possible in real scenarios because the cameras installed for security and surveillance are not of that high resolution, and it is not possible to deploy these many high-resolution cameras across the whole city. Existing datasets do not include license plate information because of the privacy of the user, making it harder for researchers to train the models and obtain results. The spatial-temporal information is also part of the user's privacy, and it is hard to obtain that information as well, so researchers usually prefer the visual information as it is easier to get the dataset for the visual appearance and it is more practical by use existing cameras.

With the help of deep models and large datasets like VeRi and VehicleID, the field of re-id has seen a remarkable gain in accuracy. Convolutional Neural Networks (CNN) and Vision transformers Networks are being used to solve this problem. There are two main approaches for the training of the model 1) Supervised Learning and 2) Unsupervised learning. In the community of vehicle re-identification, the accuracy of the supervised model is greater than the unsupervised learning. Supervised learning is done when a model is being trained on the label (Vehicle ID, Color ID) given in the dataset. On the other hand, in unsupervised learning, pseudo labels are created using a feature extractor model and then those labels are fed to the next CNN that does the classification. The main pipeline in both cases is defined in three base steps; i) Feature Extraction ii) Feature Learning iii) Distance Matric Learning. In feature extraction, the features of a vehicle are ex-

tracted using a CNN like ResNet. All the images are sent to the feature extractor and the extractor will extract those features. In the case of supervised learning, a label is attached to the feature matrix as to the original image label. In the case of unsupervised learning, a distance matrix is calculated from those features and similar features are clustered based on the distance between those matrices.

Multiple datasets exist for vehicle re-identification; like VeRi-776, VehicleID, VehicleID Small, CityFlow, VehicleID Medium, VehicleID Large, PKU-VD, and StandfordCars, etc. but the most popular among all of these are the VeRi and VehicleID and these two are explained briefly here. VeRi dataset was published in 2016 by Xinchen Liu. VeRi consists of over 50,000 images of 776 different vehicles and these vehicles were captured by 20 cameras. Those cameras were placed in 1 km^2.

Data was captured in 24 hours which makes it diverse but scalable at the same time so that it can be used for research purposes. This real-world captured data was then labeled with various attributes, e.g., Bounding Boxes, Vehicle Type, Vehicle Company/brand, and color of the vehicle, making it possible to train a different complex model on the dataset. "Spatiotemporal information and license plate information such as bounding box of plates, plate strings, the timestamps of vehicles, and the distance between the cameras"[3]. PKU VehicleID dataset was created by the NELVT, Peking University. All the data was captured in a small city in China using multiple surveillance cameras placed at distinct locations in the city. All the data was captured in the daytime. There are 221763 images in total containing 26267 vehicles. Data were manually labeled with the information of vehicle ID and vehicle model information.

1.2 Problem Statement

"Vehicle re-identification requires the capability to predict the identity of a given vehicle, given a dataset of known associations, collected from different views and surveillance cameras"[4]. Vehicle re-identification is a ranking problem; when a query image of a vehicle is given to the model, it needs to rank the given image by comparing the similarities with the database.

Moreover, vehicle re-identification inherits two main challenges that make it hard for a model to rank the vehicle correctly. The first difficulty is that when the vehicle is captured by a camera from different viewpoints, the visual appearance of the vehicle gets changes drastically. Secondly, when there are two different vehicles with the same color and model, they have a similar visual appearance. Yi Zhou explains it like this "The subtle inter-instance discrepancy between images of different vehicles and the large intra-instance difference between images of the same vehicle make the matching problem unsatisfactorily addressed by existing vision models"[2].

CNN models are being in computer vision and object re-identification, but they are not addressing the problem related to re-id. CNN faces two key issues (1) CNN can't exploit the global structural patterns and that is essential for object re-id (2) Fine-grained features are lost during the pooling and convolutions. To tackle these issues vision tranSolutions were introduced.

1.3 Solution Statement

To tackle the shortcoming of the CNNs, our proposed solution is the vision transformers. As a CNN cannot exploit the global structural patterns, it focuses on the small region because receptive fields follow Gaussian distribution. Attention modules have been introduced to exploit the long-term dependency, but they are hidden in deep layers, so they still focus on small areas, making it hard to focus on diversified discriminative parts. A transformer solves this problem by the usage of a multi-head attention module which captures the long-range dependencies that can drive a model to learn diverse vehicle viewpoints.

Fine and detailed features are important while training a model. When a feature is down-sampled in a CNN using pooling and stride convolutions, the model losses important information as the spatial resolution of the feature map gets reduced which hinders the learning of a model by reducing the discrimination ability. On the other hand, in the transformer, there is no down-sampling which helps in retaining the fine and detailed features and helps the model to learn in a better

way. The transformer's advantages over the CNN were the main reason to use Vision Transformers for vehicle re-identification.

1.4 Thesis Contributions

The contributions of this thesis have been discussed in this section.

1. The main contribution is the changing of the model in the existing pipeline.
2. Backbone was changed from Vision Transformer[5] to an enhanced form of ViT that uses a shifted window[6] method to increase accuracy more than its's predecessor and takes less time to train
3. The whole pipeline of TransReID was retrained and better accuracy was observed with the usage of a new backbone.
4. It has been evaluated and proved in this thesis that our method can be used as the backbone and is a good replacement for conventional CNNs because ViT can keep track of long-term dependencies.

1.5 Thesis Outline

The following sections will be explained in the thesis.

1.5.1 Chapter:2 Literature Review

Previous work related to re-identification has been reviewed in this chapter which includes literature from both person and vehicle re-identification. The major part includes top conference papers on vehicle re-id, deep learning-based networks, and methodologies that helped in maturing the field. Moreover, this chapter also highlights the state-of-the-art algorithms and major contributions toward vehicle re-identification.

1.5.2 Chapter:3 Datasets

There exist multiple datasets for vehicle re-identification that are being used to evaluate the various proposed model. VeRi-776, VehcileID, VehcileID small, VehcileX, CityFlow VeRi Wild are among the most commonly used datasets, details of all available datasets will be discussed in this chapter. Moreover, every dataset is not suitable for every proposed method so the datasets being used in this thesis have also been mentioned.

1.5.3 Chapter:4 Design and Methodology

In this chapter, the proposed methodology has been discussed in detail by manifesting the image and model diagrams. Each module of 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.

1.5.4 Chapter:5 Experiments and Results

The results of all tests performed using our model have been discussed in this chapter. Datasets used for 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.

1.5.5 Chapter:6 Conclusions

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.

Chapter 2

Literature Review

2.1 Vehicle Re-Identification

The world is moving towards intelligent systems with the help of deep learning algorithms. Computer vision using CNN and Vision Transformers is playing a vital part in the making of those systems. Smart and connected devices are everywhere, collecting tons of data. That data can be used to train models that can help with different problems. Current intelligent transport systems use video surveillance for smart traffic management and security. As the research related to vehicles got mature e.g., vehicle classification and detection, the researcher started working on vehicle re-identification. "There are mainly five types of deep learning-based methods designed for vehicle re-identification, i.e., methods based on local features, methods based on representation learning, methods based on metric learning, methods based on unsupervised learning, and methods based on attention mechanism"[7]. Being an important topic in computer vision, vehicle re-identification got a lot of attention in the academic community, the goal of vehicle re-id is to classify the same vehicle from the different cameras that are non-overlapping.

Identifying a vehicle from different non-overlapping cameras is challenging because of the significant difference in intra-class. The large intra-class difference means the same vehicle may look different in various scenarios; when a vehi-

cle passes through different illuminations, the visual patterns of the vehicle will change drastically and when a vehicle is being looked at from different viewpoints, the variation in the viewpoint is considerably high. "Small inter-class similarity is in reflected images of different cars may look remarkably similar. Vehicles produced by the same or different manufacturers can have similar colors and shapes, so that visual differences between two vehicle images are often subtle, making it difficult to distinguish whether the two images belong to the same vehicle"[7].

Some traditional machine methods were used at the start and hand-crafted features were used for the model training where parameters for the feature were manually adjusted. These hand-crafted feature methods include SIFT, HOG, and LBP. SIFT extracts local key features that are invariant to illumination, size of object, and rotation but at the same they require high computation so cannot be used in real-time systems and feature extraction of SIFT is very low on smooth edges. HOG can extract edge information in a better way even in the presence of different lighting conditions and colors. As it focuses on edges, not on the local key features, the amount of calculation require is less than SIFT but at the same time, it is sensitive to occlusion and noise. LBP focuses on texture information which makes it invariant to light. The calculations in the LBP are not complex which makes operations faster, on the other hand, LBP doesn't perform well when the directions of textures are changed[8].

With the introduction of deep convolutional neural networks (CNNs), the hand-crafted features were abandoned and CNNs with deep hidden layers became so famous that they were being used in every field of computer vision, such as object classification, detection, image semantic segmentation, object tracking, etc. and they were achieving some serious accuracies. Due to this, the researchers started to tackle vehicle re-identification using CNNs

2.2 Vehicle Re-Identification using Local Features

At the start of vehicle re-identification using deep learning, researchers were focused on global features of the global features which caused a bottleneck in ac-

curacy, so they started to pay attention to the local feature instead as in-vehicle differences usually appear in the local features. Region segmentation and key-point location are the most wildly used methods for the extraction of local features.

Wang et al. [9] used both region segmentation and location of key-point and to get segmentation results of different regions by marking twenty different points on the target vehicle. The convolutional neural network was extract local features from those marked key points and then those key points were fused with the global features to get the final appearance feature vector of a target vehicle. Then, those fused vectors were used to solve the vehicle re-identification problem; as those features can directly be compared for query image features and features with other image features in the database. The proposed solution was a good addition to the vehicle re-identification community, but it was limited by the dataset diversity; data should not be limited to some viewpoints of the vehicle, it should have multiple viewpoints for each vehicle. In a real scenario, it is hard to get a perfect dataset like that as there would be too many images for each vehicle making the dataset too big and it is not possible to collect images of vehicles with every viewpoint. Moreover, for every viewpoint, the key points should be labeled and if there are various viewpoints then there would be too many images to label. "In addition, key points need to be labeled for different angles of the vehicle image on the collected dataset, so the number of key points that need to be labeled is large which results in a huge workload. Therefore, the method was complicated in terms of feasibility and workload" [7].

Local features are important for vehicle re-identification because it is not possible to differentiate between the vehicles using global features only, therefore some researchers used both local and global features for vehicle re-id which led to improvements in model accuracy. Liu et al. [3] outlined a deep model named as Region-Aware deep model(RAM) that extracted both local and global features, those features were combined to get more discriminative details of the vehicle as local features can provide details of the vehicle. Besides, they have used vehicle color, id, and type information to future improve the ability of the model to discriminate between the vehicles. A similar approach was also introduced by He et

al. [10] by training a model end-to-end by using the local and global features but they added a detection branch as well. They used local features of various parts of the vehicle such as front, back, sides, windows, lights, and brand e.g., KIA, etc. Then they used the part attention mechanism to formalize the global module, by these formalized global features, model was able to differentiate between different vehicle accurately. Peng et al.[11] introduced a method called Multi-Region Model which uses multiple regions to extract features and a Spatial transformer model was trained for each region that helps in the localization of the features. Then they used a context ranking algorithm, the algorithm could rank the different vehicles based on context and content which further increased the accuracy of the vehicle re-identification.

Ma et al.[12] came up with a model that could learn feature embeddings efficiently, the model was based on Grid Spatial Transformers Network (GSTN), and the model was able to divide local features from the global features and locate the vehicle. "Besides, residual attention was conducted to give an additional refinement for a fine-grained identification, the refined part features were fused to form an efficient feature embedding finally, so that improved the accuracy of vehicle re-identification. In summary, the advantages of methods based on local features are reflected in it can capture unique visual clues conveyed by local areas and improve the perception of nuance, which helps a lot to distinguish between different vehicles and improve the accuracy of vehicle re-identification. Besides, many researchers combine local features with global features to improve the accuracy of vehicle re-identification. Hoverer, the disadvantage of methods based on local features is the extraction of local features will significantly increase the computational burden"[7].

2.3 Vehicle Re-Identification using Representation Learning

Multiple camera angles exist in the real-life application of vehicle re-identification, obtain images from those camera means getting different viewpoints that can re-

sult in bad differences in the local key feature areas. Only using local features, is not possible to achieve high accuracy. As the CNNs got developed, the researcher made serious progress in representation learning. Features/representations are learned by the transformation of input data by using convolutional neural networks that can further be used for different computer vision tasks such as prediction, detection, and classification. Convolutional neural networks are trained on a big dataset like ImageNet; containing millions of images so that they can learn representation automatically. Representation learning has been applied to person re-identification and has given some serious results, so researchers started to apply it to vehicle re-identification as well.

Learning discriminative features are important to learn from the different vehicle viewpoints, Zheng et al.[13] proposed a deep convolutional network to learn the features embedded with other vehicle attributes, including color, type, and camera view, the solution was named DF-CVTC. With the combination of these attributes the features were improved by a large margin, once the features were learned by the model, another model VS-GAN was trained to learn these representations and enhance the variance of the data. Huang et al.[14] proposed a deep feature fusion with Multiple Granularity (DFFMG) that used a combination of global and local features by using two directions (i.e. vertical and horizontal). "DFFMG consisted of one branch for global feature representations, two for vertical local features representations and other two for horizontal local features representations"[14].

Different representation learning-based methods proposed some unique and novel ideas that helped in getting good accuracies. Hou et al. [15] introduced proposed a deep representation learning based that can get random occlusion features for vehicle reidentification algorithm. The algorithm randomly occluded the training images which was almost close to a real-life situation where some parts of vehicles are occluded, this way they were able to increase the dataset images that help in avoiding the overfitting of the model. Features extracted from both occluded and normal images were used to train the model and the results were promising. Various viewpoints of a vehicle are interlinked and to exploit the relationship between these views Alfasly et al.[16] proposed a Long Short-Term Memory (LSTM) based

solution. The framework could learn the relationship between multiple viewpoints of the vehicle using the representation of images which helped in the improvement of the accuracy of the model. They used Kullback-Leibler divergence which can improve the performance of the model for vehicle re-identification and other tasks as well such as detection and classification.

To solve the problem of data labeling on image datasets and unlabeled videos, Wu, et al. proposed a CNN-based model that could automatically extract the representations based on the space-time labels and label those in positive and negative samples. The feature extractor was trained on multi-label data and then it was finetuned on the vehicle dataset so that model could learn the features of the vehicle and perform better on test videos. To help the model in learning a new distance loss was proposed that work in a way that it creates sets for the same vehicle and add similar image to one set and different image to different sets. Then they trained the mode on those sets in a way that the distance between images similar was optimized and this proposed loss performed better than triplet loss. To learn discriminative features/representation, Jiang et al.[17] proposed a vehicle reidentification model that was multi-attribute driven, essentially making the model, a multi-branch model with an enhanced re-ranking scheme. They added vehicle attributes such as color and model to increase the discrimination ability of the multi-branch model. They also used special-temporal information to divide the representations of vehicles into sets then Jaccard similarity is used to find the distance between the images of different sets.

Two types of vehicle re-identification streams are being carried out now. The first one is to consider the task as a classification problem and train the deep learning model for vehicle re-identification in a supervised manner, by feeding a lot of images labeled data to the model and using various loss functions for classification to optimize the model. The loss is computed while training when the model predicts a vehicle ID, it is compared with the original label and then using multiple iterations of back-propagation and forward steps the loss is optimized. Nevertheless, vehicle models that are being produced every year with new shapes and colors are large in numbers, and solving this problem with simple classification

will lead to the over-fitting in the vehicle re-identification domain because when there is a large amount of data, it becomes difficult to classify different vehicles that can drive the models towards bottleneck in accuracy. The other technique finds the vehicle re-identification as a vehicle verification problem, in verification, IDs are assigned to the two pictures of the vehicle and then verification loss is optimized between two images. Having said that, verification learning takes too long as it is a one-to-one comparison between a pair of images, it can only compute loss between images in pair. Moreover, the generalization ability of the model suffers as it only uses vehicle ID and does not make use of other attributes of the vehicle such as type, color, and model.

In a nutshell, vehicle re-identification is aided by representation learning, the feature extractor models can extract the representation automatically which helps in the elimination of hand-crafted features, and they are robots in a different situation. Training a representation model is stable and results can easily be reproduced despite those methods based on representation learning are not good enough in terms of generalization and as they can be overfitted on the large dataset.

2.4 Vehicle Re-Identification using Metric Learning

Metric learning[18] is the method of converting images into feature space using feature transformation then in the feature space cluster are formed for similar vehicles. Metric learning is widely used for vehicle re-identification, face recognition, and person re-identification. In metric learning, the model is trained in such a way that the distance between similar vehicles is reduced and the distance between different vehicles is increased; the similarity between similar vehicles is high and the between different vehicles is low. “Therefore, metric learning requires certain key features of the learning objectives, that is, individualized features. When distinguishing different vehicles, the appearance characteristics of the vehicles are very similar, these features belong to the common features between vehicles. Distinguishing features like the paint, stickers, scratch marks on the vehicle, the annual inspection position of the vehicle on the front windshield, decoration, and

tissue boxes are used to distinguish the different characteristics of the two cars. Metric learning distinguishes different identities by learning key distinguishing features"[7].

Contrastive loss, quaternion loss, and triple loss are the most widely used losses in metric learning. Given any picture x, the feature vectors fx for the image can be extracted using forward propagation then the distance between extracted features is calculated using the Euclidean distance formula.

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{2.4.1}$$

Siamese networks are also popular in-vehicle re-identification, Zakria et al.[19] brought in a vehicle re-identification technique in which a vehicle is selected from the gallery then the selected vehicle is verified using a license plate. The model learns the global features vector of the vehicle, the local features are also learned using the local region channel and they also incorporated the other features like type, model, and color. In the end, they used Siamese neural network to check the accuracy of model prediction(re-identification). Liu et al.[20] created a method that looks for the vehicle in a step-by-step fashion. A classifier learns the visual features such as color, type, and vehicle model, the Siamese networks match the license plate and use spatial-temporal information to classify the vehicle.

Combing Siamese Networks with other models resulted in good accuracies. Shen et al.[21] created a combined framework using the Siamese network and the Path-LSTM model. These streams work in a way that the first branch Siamese network is used to extract visual similarity while the other branch consists of LSTM to learn spatial-temporal data as LSTM know for working best on time series data, at the end, results from both branches are merged to get the re-identification results. A deep learning-based two-branch Multi-DNN Fusion Siamese Neural Network was proposed by Cui et al.[22] to tackle the problem of vehicle re-identification. They placed some random marks on the windshield and use those marks as an attribute alongside the other attributes such as color and vehicle model, then they mapped those attributes into a Euclidean space where the distance can be calcu-

lated between two similar of different vehicles based on the Euclidean distance, that distance is called the similarity between two vehicles.

Inspired by the triplet loss, Liu et al.[20] created a Deep Relative Distance Learning (DRDL) method that consists of two branches deep convolutional neural network to show the image of a vehicle to the Euclidean space. In Euclidean space, the distance between two random vehicles can be used to calculate the similarity between vehicles. Group-Sensitive-Triplet-Embedding (GS-TRE) method was introduced by Bai et al in which each vehicle was considered as a group. The number of groups was dependent on the number of classes in the dataset, similar vehicles with the same ID were sent to the same group, this way the intra-class difference got easier to learn. Kumar et al.[23] proposed a baseline model for vehicle re-identification that utilizes the triple embeddings with a combined loss of both triple loss and contrastive loss. Zhang et al.[24] proposed a similar approach with the help of Triplet loss, the model was trained on a triplet of an input image with both positive and negative samples from the database to calculate the similarity between the positive samples. They used a classification model with the help of triplet loss that enhanced the classification ability of the model up to a large extent.

Moreover, A network named Deep-Joint-Discriminative-Learning (DJDL) was proposed by Li et al. [25]. Extraction of discriminative feature representations from vehicle images was the main goal of the network. DJDL consisted of four subnetworks, identification, recognition, verification, and triplet network, each adding its functionality to the final output. Triplet makes sure that the distance between similar samples is minimum, verification keeps the relation between both positive and negative samples, and identification and recognition are responsible for making use of different properties of the image. All these subnetworks were optimized for vehicle re-identification and the results were promising. Wang et al. explained the limitations of the classification in supervised learning like this “For supervised learning, the category is usually fixed so that the SoftMax cross-entropy loss function can be used to train to meet the classification requirements. But sometimes, the category is a variable, especially for vehicles, the variety of models is different

and will be updated or the quantity will change at any time. The trained classification model has poor generalization ability or is prone to over-fitting so vehicle re-identification tasks are not well done with only use classification learning, using triplet loss can solve such problems"[7].

Triplet loss can get an advantage of the model when there is detailed variation in the data that triplet can better learn the details and different characteristics between input samples. While training the common feature of the vehicle are not given attention but the attention is given to the feature that can differentiate two samples like scratches, paint writings, windscreen, different stickers on the windscreen, tissue boxes, accessories, decorative items, etc. these features can help in getting better accuracy in vehicle re-identification as the other features such as color, shape, bagging of the company can be similar on a vehicle that can confuse the model.

2.5 Vehicle Re-Identification using Unsupervised Learning

Labeled data plays a vital role in every field of deep learning, supervised learning requires a lot of labeled data. To label the raw data, a lot of effort is required which is not always feasible. Unsupervised learning techniques do not require labeled data, but they can infer directly from unlabeled data using some clustering algorithms, and these algorithms have successfully been applied to various re-identification tasks.

Deng et al. proposed an Image-to-Image-Cross-Domain-Adaption unsupervised technique by using self-similarity and dissimilarity using GAN that preserves the similarity between vehicles and used contrastive loss for re-identification. Bashir et al.[26] presented a technique "that essentially formulates the whole vehicle re-ID problem into an unsupervised learning paradigm using a progressive two-step cascaded framework. It combines a CNN architecture for feature extraction and an unsupervised technique to enable self-paced progressive learning. It also incorpo-

rates contextual information into the proposed progressive framework that significantly improves the convergence of the learned algorithm. Moreover, the approach is generic and has been the first attempt to tackle the vehicle re-ID problem in an unsupervised manner. The performance of the proposed algorithm has been thoroughly analyzed over two large publicly available benchmark datasets VeRi and VehicleID for vehicle re-ID using image-to-image and cross-camera search strategies and achieved better performance in comparison to current state-of-the-art approaches using standard evaluation metrics"[26]. Marín-Reyes et al.[4] applied a method that creates the annotations from the video in an unsupervised way, that can be used to train the model. Bashir et al.[27] trained a model with self-progressive learning architecture, the technique deeply learned the representation of unlabeled data.

Goodfellow et al.[28] presented a framework GAN that contains two networks; generator and discriminator. The generator is given the samples with added noise and asked to generate the real sample from the given data while the discriminator keeps check of whether the generated data is good enough or not. Generative Adversarial networks are being used in various deep learning tasks such as image generation, the real reason behind the success of the GAN is that it can create synthetic data which is indistinguishable from real data. Besides, there exist many iterations of GANS such as AC-GAN, Info-GAN, and Cycle-GAN that tend to improve the performance of originally presented GAN with certain conditions. By exploiting, the synthetic data generation ability of GAN Zhou et al.[2] proposed a viewpoint-aware attentive multi-view inference (VAMI) model to solve the problem of variations in viewpoint by using visual information only. VAMI uses a single viewpoint on of vehicle and transforms that view into a global viewpoint using GAN in such a way that pair-wise distance can be learned. Given a query image to the generator model, it generated all the viewpoints of a vehicle, and the discriminator makes sure that the created viewpoints are perfect so that features can be better learned, this approach got promising results.

"In summary, unsupervised learning technology can make use of unmarked input data to improve generalization ability. Among the vehicle re-identification

methods based on unsupervised technology, GAN-based methods are widely used. GANs can generate multiple perspective features for a single perspective image and use the feature to solve the vehicle re-identification problem under multiple viewing angles, in addition, GANs can be used for image-to-image translation to better solve the problem of inconsistent distribution of different data domains. But using GANs for image generation needs to overcome the problem of difficulty in convergence, and balance the two models in training, thereby avoid unstable training situations"[7].

Chapter 3

Datasets

3.1 Datasets for Vechilce Re-Identification

There exist multiple benchmark datasets for vehicle re-identification such as VehicleID, VeRi-776, CityFlow, Veri-Wild, VehicleX, VRAI, Vehicle-Rear, Vehicle-1M, CityFlow, PKU-VD, and VehicleID small, etc. VehicleID, VeRi, VeRi Wild, and CityFlow are among the most used datasets for state-of-the-art model training in the community of vehicle re-identification. In this chapter, all datasets will be described in detail.

3.1.1 VeRI

The VeRi dataset was proposed by Liu et al.[3] in 2018 with a deep learning technique for vehicle re-identification. The images of the dataset were extracted by the real-world surveillance cameras, 20 cameras were placed at different viewpoints; covering a radius of $1\text{km}^2 for 24 hours$.

50k images were extracted from the feed containing 776 vehicles, making the dataset feasible for vehicle re-identification and detection task. Captured images were labeled manually with various attributes such as BBox, color, type, and vehicle brand. Each was captured in various illumination, viewpoint, resolution, and occlusion conditions by 2 18 cameras. The data is also labeled with spatial-temporal information such as license pale, license plate number and timestamps

of vehicles, and distance between the cameras.

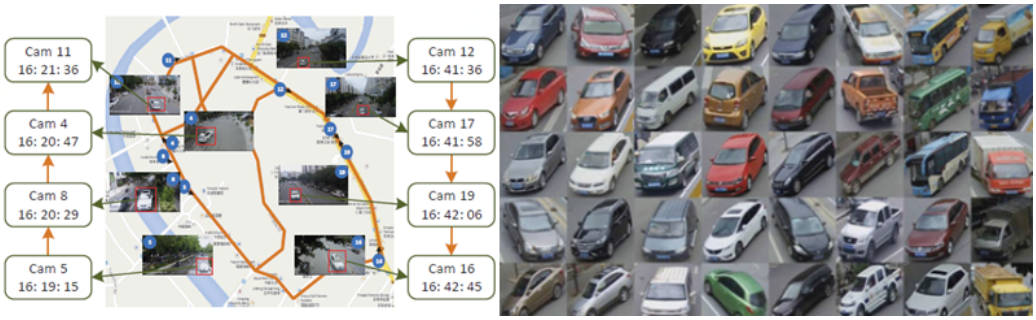

Figure 3.1: The position of different cameras in the VeRi dataset has been shown in the figure, moreover different vehicles from the dataset are also shown.[3]

3.1.2 VehcileID

The PKU VehicleID dataset was created by Peking University, dataset contains 26267 unique vehicles with 221,763 images in total, the dataset was collected in a small city in China from real surveillance cameras. Each image is given an ID that corresponds to the real vehicle; besides, 10319 vehicles were manually labeled with the model information.

3.1.3 CityFlow

CityFlow was made using 3 hours of HD video from 40 different cameras that place throughout the whole of 10 intersections in a city where traffic was flowing rapidly, the longest distance between two cameras was around 2.5KM. The dataset contains around was annotated for vehicle ID, view angle, model, and the condition of the traffic flow by drawing 200K bounding boxes. Besides, spatial-temporal information such as Camera calibration and geometry information was added so that it can later be used in methods based on spatial-temporal information.

3.1.4 Vehicle-1M

The vehicle-1M dataset contains around 936,051 images of 55,527 vehicles belonging to 400 different vehicle models. All these images were collected from the rear

or front of the vehicle in all possible lighting scenarios from day to night. All the vehicles are labeled with an ID that denotes their identity in the real world as well (i.e "Honda-Civic-2013") which indicates the brand, model, and year of the model.

3.1.5 VRAI (Vehicle Re-Identification for Aerial Image)

For UAV-based applications, VARI was proposed as a large-scale dataset for vehicle re-id. The dataset consists of 137,613 aerial images of 13,022 unique vehicles.

3.1.6 VeRi-Wild

VeRi-Wild was known to be the largest vehicle re-id dataset according to CVPR 2019. The dataset consists of 416,314 images of 40,671 vehicles and these images were captured from 174 cameras. The feed was collected in a month for 24 hours and it was coming from CCTV cameras.

3.1.7 Vehicle-Rear

As the name indicates, this dataset contains rear images of nearly 3000 vehicles, all the images were obtained from a 3-hour-long feed of high-resolution cameras. Annotation of this dataset includes model, manufacturer, color, year, position, and license plate information of the vehicle.

3.1.8 VehicleX

As dataset collection is a challenging task, researchers have come across various methods of making a dataset. VehicleX is an example of such an effort, it is a synthetic dataset that was created in Unity (A game development tool). It consists of 1362 uniquely created vehicles with the ability to edit any attribute.

VeRi-776 is known to be the most used dataset for various vehicle re-identification benchmark models so we have used this dataset for model evaluation.It contains

50000 images of 776 unique vehicles with the information of cameras which is essential for our proposed method.

Name	Images	Vehicles	Cameras	Real
VeRi-776	50,000	776	20	Yes
VehcileID	221,763	26267	-	Yes
CityFlow	20,0000	-	40	Yes
Vehicle-1M	936,051	55,527	-	Yes
VRAI	137,613	13,022	-	Yes
VeRi-Wild	416,314	40,671	174	Yes
Vehicle-Rear	3 Hours Video	3000	-	Yes
VehicleX	-	1362	-	Synthetic

Table 3.1: Details of all dataset avaible for vehicle re-identification is shown in this table, name, number of image, number of vehicles, number of cameras, and state of the dataset is shown..

Chapter 4

Design and Methodology

4.1 Problem Definition

To address the problem of vehicle re-identification, a novel approach based on Vision Transformers that have recently been introduced by the computer vision community. Transformers can safely replace the conventional convolutional neural networks; in some scenarios, they can even beat the accuracy of CNN's because they can leverage from keeping the long-term dependencies intact while the CNN can only keep local neighborhood information and loses information when it performs pooling and convolutions.

The first effort to use a vision transformer for vehicle re-identification was proposed in a paper by TransReID[29]. They used code from Reid-Strong-Baseline[30], and PyTorch-Image-Models and they got some competitive results. The research gap that existed in the technique is that there are some improved vision transformer models such as Swin Transformer. We have changed the model from Vision transformer to Swin Transformer and the results were promising.

4.2 Architecture Design

The main architecture is consisting of two main modules Swin Transformer and TransReID, both have been used to get results. The main module used for the

baseline is TransReID, which was originally using Vision Transformer (ViT), we have replaced the model with a Swin transformer[31]; uses a shifting window scheme to get better performance than ViT.

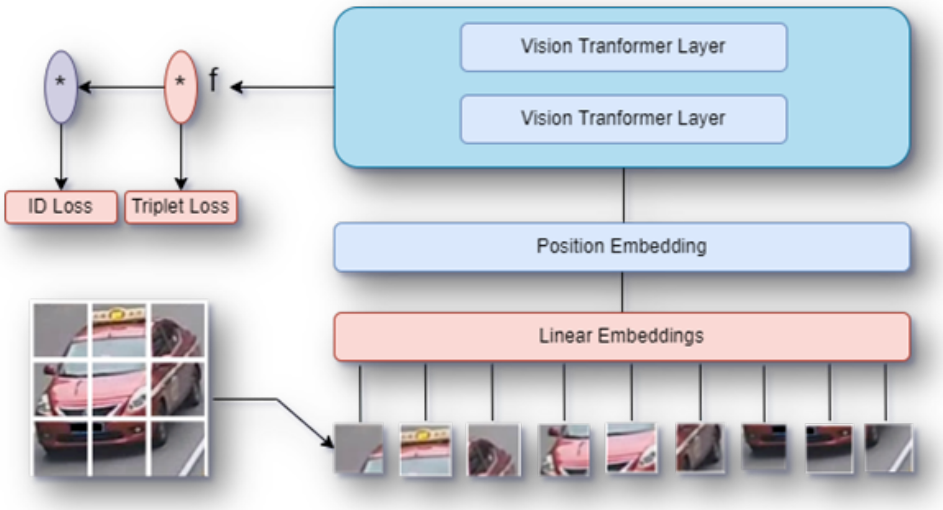

Figure 4.1: This figure explains the architecture of the proposed methodology, which uses the Swin Transformer as the backbone. We have used two types of losses ID loss and Triplet Loss.

4.2.1 Swin Transformers

Swin transformers are an enhanced form of vision transformers that use shifted window scheme. A patch Splitting module splits the input RBG image into non-overlapping patches. Each patch is named a "token," and its raw values are concatenated with the pixel values. In implementation[6], a path size of 4*4 was used, which result in a feature dimension of 4*4*3 = 48. A linear embedding layer projects those features maps to an arbitrary dimension (Denoted as C). These patches go through multiple "Swin Transformer blocks," and they get modified by self-attention. In Stage One, the linear embeddings and number of tokens are maintained by the transformer block.

As the network gets deeper, the number of tokens is reduced by the path merging model which produces a hierarchical representation. Feature of each group

of 2*2 are merged by the first path merging module and then the linear layer concatenated the 4C-dimensional features, by merging, the tokens are reduced by a factor of 2*2=4 (2*down-sampling of resolution), making output dimension 2C. By keeping the resolution of features to W/8*W/8, Swin transformer blocks are applied after the feature transformation. Stage 2 is regarded as the first block of path merging and feature transformation. The output resolution is doubled after every block, making W/8*W/8 to W/16*W/16 and W/16*W/16 to W/32*W/32 as the above step is repeated two in stages 3 and 4. Hieratical representations are produced after these steps which makes it like the feature dimension produced by the convolutional neural network, e.g., ResNet and VGG. As the dimension of the feature map is similar, Swin Transformer can be used as a backbone for different vision tasks such as re-identification, detection, and classification.

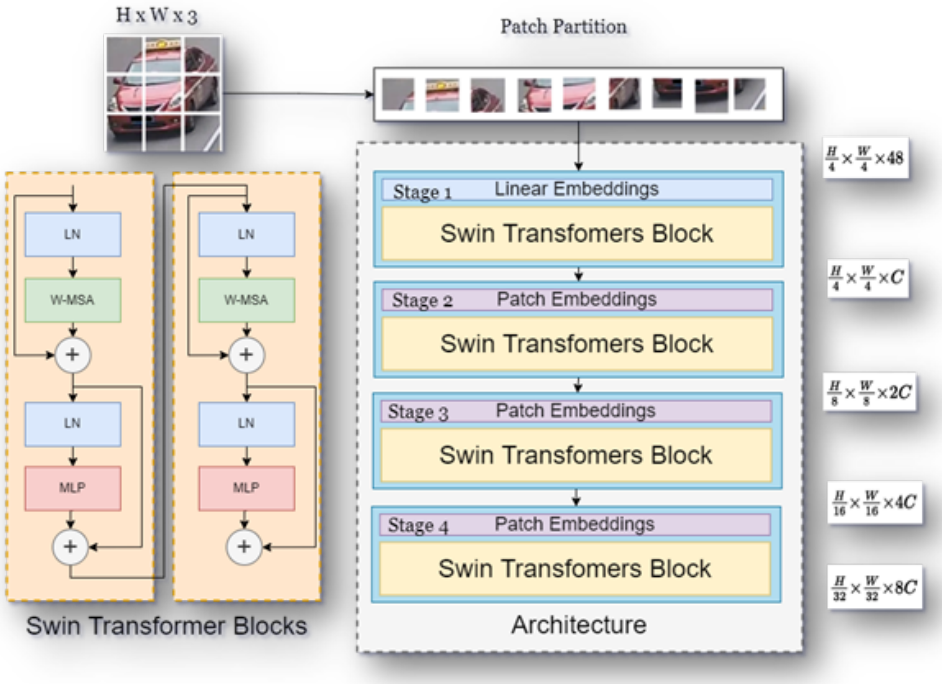

Figure 4.2: This figure explains the architecture of the Swin Transformer, that have 4 transformer blocks and a match embedding module.

Swin Transformer Block : Swin Transformer is based on Vision Transformer (ViT), the multi-head self-attention (MSA) module is replaced by a module by a shift window, and other blocks are kept the same. The architecture of Swin Transformers is shown in Figure 3, "a Swin Transformer block consists of a shifted window based MSA module, followed by a 2-layer MLP with GELU nonlinearity in between. A Layer Normalization (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module"[31].

Shifted Window-based Self-Attention : Originally proposed Vision Transformer and ViT for image classification both use the global self-attention; the association between all tokens is learned by the self-attention module. The complexity of these computations is "quadratic" according to the number of tokens, meaning that the ViT requires a lot of tokens for better prediction. The number of tokens is directly proportional to the number of images and the resolution of images which makes it unsuitable for various computer vision problems.

Swin Transformer solves this problem using "Self-attention in non-overlapping windows" and "Shifted window partitioning in successive blocks". In Swin transformers, self-attention is computed within the local windows. The image is divided into equally partitioned non-overlapping windows. Each window contains M x M patches then the multi-self-attention module and a window, the complexity for h x w patches can be calculated using the equation below.

$$MSA = 4hwC^2 + 2(hw)^2C \tag{4.2.1}$$

$$W - MSA = 4hwC^2 + 2M^2hwC \tag{4.2.2}$$

"Where the former is quadratic to patch number hw, and the latter is linear when M is fixed (set to 7 by default). Global self-attention computation is generally unaffordable for a large hw, while the window-based self-attention is scalable"[31].

There is no connection between the windows in the window-based self-attention module which reduces the modeling power of the window self-attention module.

To add the connection/relation across the windows, a shifted window approach was introduced that not only switches between two consecutive transformer blocks but also maintains the efficient computation of non-overlapping windows.

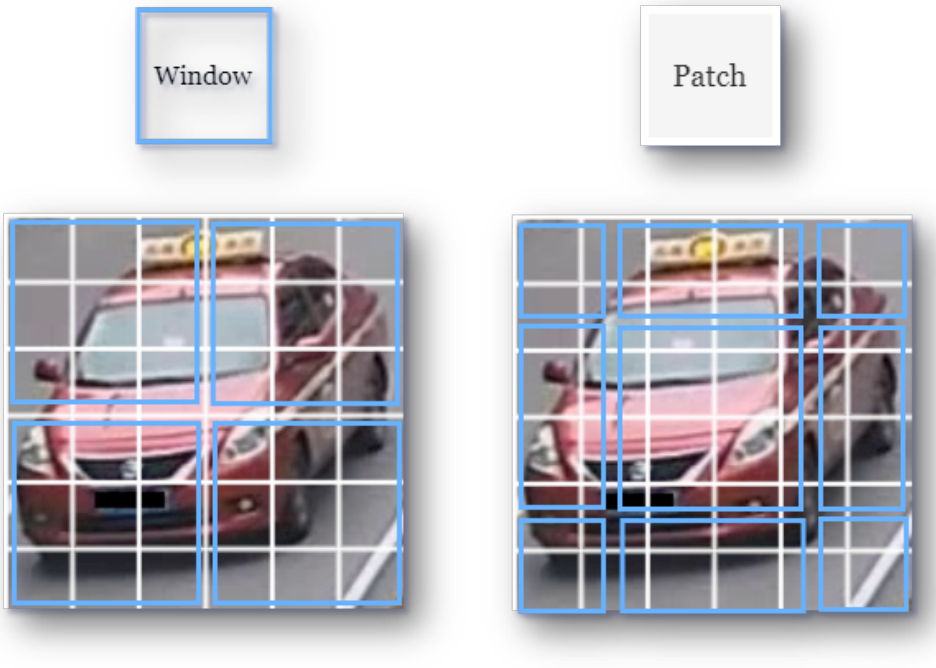

Figure 4.3: As illustrated in Figure 4, in the part a regular window splitting approach is used, it starts from the top left corner and maps 8 x 8 feature to 2 x 2 window with size 4 x 4 where M=4. "Then, the next module adopts a windowing configuration that is shifted from that of the preceding layer, by displacing the windows by ([M/2] , [M/2]) pixels from the regularly partitioned windows"[31]

4.2.2 TransReID: Transformer-based Object Re-Identification

This paper has been used as a base paper for this thesis. TransReID is based on transformer-based image classification but with an improved feature extraction method. To improve the features two novel modules have been added; Jigsaw and Side information embeddings (SIE), using both modules, the models jointly is trained end-to-end.

For sake of object re-identification, a transformer-based baseline has been in-

troduced. The method includes two stages; feature extraction and supervised learning. An image is denoted x RH x W x C, where width, height, and image channels are denoted as W, H, C, then the fixed-sized patches are made by the image. Input sequences are embedded with a new learnable token called [cls] embeddings. Global features f is represented by the input token. Position embeddings are added to fuse the spatial information then transformer layers are fed these input sequences that can ne expressed as:

$$Z_0 = [x_{cls}; F(x_p^1); F(x_p^2) \ldots; F(x_p^N)] + P \tag{4.2.3}$$

"Where Z0 represents input sequence embeddings and P R(N+1) x D is position embeddings. F is a linear projection mapping the patches to D dimensions. Moreover, transformer layers are employed to learn feature representations"[31]. The receptive field of transformers is always global which solves the problem of CNN-based limited receptive field. Moreover, in CNN details are lost while doing downsampling and convolution, it is not applicable here.

Inputs are split into non-overlapping patches in transformer-based models that result in loss of structure around the local neighborhood patches, in this paper, a sliding window was introduced to include information from neighbors as well. If Step size and patch size are denoted as S and P, respectively then the shape of the overlapped area is denoted as (P - S) x P. There will be N patches from an image with size H x W.

$$N = N_H * N_W = [\frac{H + S - P}{S}] * [\frac{W + S - P}{S}] \tag{4.2.4}$$

"Where [.] is the floor function and S is set smaller than P. NH and NW represent the numbers of splitting patches in height and width, respectively. The smaller S is, the more patches the image will be split into"[6]. It can be inferred from the equation that increasing the number of patches will increase the accuracy, but computation costs will be increased.

Position Embeddings : Position embeddings trained on ImageNet can't directly be used as image resolution in the original classifier is different from the image in the re-id tasks. To tackle this problem of having multiple resolutions, a bilinear 2D-interpolation technique was introduced. Position embedding here is also a learnable parameter like in vision transformers.

Supervised Learning : The model was optimized for global features with triplet and ID loss. The ID loss is the cross-entropy loss except for label smoothing.

Side Information Embeddings : Though the obtained features are fine-grained, they are still sensitive to variations in viewpoints which is called scene bias due to which the model might not be able to distinguish the objects in the different viewpoints. Therefore, the information of viewpoints is also embedded with the features given to the model using a module called "Side Information Embedding" (SIE), making the features invariant to different viewpoints. Side information has also been made a learnable parameter like position embedding was encoded as a learnable parameter to retain the position information, along with patch embeddings and position embeddings, SIE is also fused into transformer encoder. "In specific, suppose there are NC camera IDs in total, we initialize learnable side information embeddings as SC 2 RNCD. If the camera ID of an image is r, then its camera embeddings can be denoted as SC[r]. Different from the position embeddings which vary between patches, camera embeddings SC[r] are the same for all patches of an image. In addition, if the viewpoint of the object is available, either by a viewpoint estimation algorithm or human annotations, we can also encode the viewpoint label q as SV [q] for all patches of an image where SV 2 RNV D, and NV represents the number of viewpoint IDs"[31].

SIE module can be used to embed any sort of information that is needed to add into the model, in this paper only camera and viewpoint information was added using the module.

Chapter 5

Implementation

In this chapter, the implementation details are briefly described.

5.1 Dataset Details

The dataset being used are VeRi and VehicleID, all images are resized to 224 x 224 as the input of Swin Transformer is 224 x 224 for the base model. All the training images are augmented, and the operation performed contains; random horizontal, padding, random cropping, flipping, and random erasing. The datasets used to evaluate the proposed method include VeRi and VehicleID, the data have been split into the train, validation, and gallery image with the portion of 80, 10, and 10 percent respectively. Information on the camera is provided in the VeRi Dataset while VehicleID does not include that, details of both datasets are summarized in the table below.

Dataset	Object	Images	Classes	Cameras	Views
VeRi-776	Vehicles	50,000	776	20	8
VehcileID	Vehicles	221,763	26267	-	2

Table 5.1: This table shows the different variants of the Swin Transformer in detail. Their resolution, Ranks and Parameters have been mentioned in the table.

As discussed earlier the baseline model of the original paper has been changed from

ViT transformer to Swin Transformer because the accuracy of the Swin Transformer is better than the Vision Transformer. The Swin Transformer fixes two fundamental issues of ViT which were a bottleneck for accuracy, shifted window attention, and hierarchical feature maps; it helps to reduce the amount of data needed for training as well as feature learning is more robust.

5.2 Training Settings

After testing various parameter settings, the following settings were finalized because we got maximum accuracy using these particular settings. The batch size of the model was set to 64 by using 4 images of each class, Stochastic Gradient Descent (SDG) optimizer was used with a momentum of 0.9 having a weight decay of 1e-4. Using cosine learning rate decay the learning rate is initialized as 0.008. If not specified, the value of m = and k = 4 for vehicle Re-identification datasets. All the experiments were performed on Nvidia Tesla p40 with 24Gb of GPU-Memory with PyTorch toolbox.

5.3 Model Details

Weights used for initialization are from the provided by the official GitHub repository of Swin Transformer that is trained on ImageNet-1k.The model used in our proposed method was the first in the table "Swin-B," with 81.2 percent Rank-1 accuracy on the ImageNet-1k dataset, having 88M parameters.

5.4 Hardware and Runtime

Hardware and runtime are important factors in the deep learning community. We have trained our model on an NVIDIA Tesla P40 that is based on NVIDIA Pascal architecture, it has 3840 CUDA Cores, 24GB GDDR5, and works on 250W. Our model was trained on VeRi data with a batch size of 64 and for 120 epochs, the time taken was 36 hours. The model was trained multiple times to get the average

Name	Resolution	Rank-1	Rank-5	Params	Used
Swin-T	224x224	81.2	95.5	28M	NO
Swin-S	224x224	83.2	96.2	28M	NO
Swin-B	224x224	83.5	96.5	28M	YES
Swin-B	384x384	84.5	97.0	28M	NO

Table 5.2: Details two main datasets VeRi and VehicleID for vehicle re-identification is shown in this table. Besides, name, number of image, number of vehicles, number of cameras, and state of the dataset is manifested.

accuracy but the time for training remains the same with negligible differences.

Chapter 6

Results and Discussion

6.1 Data

In this section, we are going to discuss the results of the proposed method on the used dataset. We have used a benchmark dataset to evaluate our model, VeRi-776 was the dataset used for the evaluation of the proposed model. VeRi-776 was proposed by Peking University, China. They proposed a paper with a baseline for vehicle re-identification and VeRi-776 was introduced in that paper. VeRi-776 was made by using real footage of traffic in China, 20 cameras were placed within a radius of 1km, all cameras were placed at different angles, and the feed coming from the cameras was non-overlapping. Cameras recorded the footage for 20 hours and then all the videos were converted into images. Images were then labeled by the ID of the vehicle with other attributes such as; shape, color, license plate, etc. Overall, 50,000 images were obtained with 776 unique vehicles. This dataset was used for benchmarking, and we followed the standard Train/Val/Test splits and data.

6.2 Evaluation Protocols

To evaluate the performance of the proposed network for vehicle re-identification, the convention in the re-identification community was followed, we have used mean

Average Precision (MAP) and Cumulative Matching Characteristic (CMC).

mAP : to calculate the accuracy of the different models for object detection this is popular matric, the average precision for recall value in range 0 to1 is computed in average precision. Finding the area under the precision-recall curve is called average precision. Besides, mAP is a 101-point interpolated definition of average precision. Average precision is the average of all categories that are traditionally called "Mean Average Precision."

Rank : we have used rank as a metric to compute the accuracy of the model, rank is the ability of a model to distinguish the desired object from the gallery of images. For example, Rank-5 with 90 percent accuracy means the model is 90 percent that the object in one of 5 images, similarly, Rank-10 is about the confidence of the model in 10 images. The higher the rank value, the more accurate the model will be, and vice versa.

6.3 Training Settings

Our proposed model was trained on a single Nvidia Tesla p40 GPU, training was done for 120 epochs and the batch size of the model was set to 64 by using 4 images of each class, Stochastic Gradient Descent (SDG) optimizer was used with a momentum of 0.9 having weight decay of 1e-4. Using cosine learning rate decay the learning rate is initialized as 0.008. If not specified, the value of m = and k = 4 for vehicle Re-identification datasets. All the experiments were performed on Nvidia Tesla p40 with 24Gb of GPU-Memory with PyTorch toolbox. Pretrained weights used for initialization are from the provided by the official GitHub repository of Swin Transformer that was trained on ImageNet-1k.

6.4 Evaluation on VeRi-776

VeRi-776 has been used as a benchmark dataset to evaluate the performance of the proposed model. For background, we have replaced the Vision Transformer model with the Swin Transformer model as it is an improved version of ViT and has better accuracy on various benchmarks, it outperformed models like ResNet, ViT, and DeiT. The details are shown in Table below.

Model Name	Image Size	Parameters	Flops	Top-1 Acc
R-101x3	384x384	388M	204.6G	84.4
R-152x4	480x480	937M	840.5G	85.4
ViT-B/16	384x384	86M	55.4G	84.0
ViT-L/16	384x384	307M	190.7G	85.2
Swin-T	224x224	29M	4.5G	81.5
Swin-B	224x224	88M	15.4G	85.2
Swin-B	224x224	88M	47.0G	86.4
Swin-L	384x384	197M	103.9G	87.3

Table 6.1: This table shows the comparison of the accuracy of the Swin Transformer model with various benchmark CNN-based networks and the originally proposed Vision Transformer.

As shown in the table Swin Transformer has fewer parameters but higher accuracy which means the model will need less time to train as there are fewer neurons that are needed to be finetuned.

Now, we will discuss the performance of our proposed network on the VeRi dataset. The performance of the model was up to the mark as it matches the accuracy of the base paper model with the base model discussed in the paper which proves that if the Swin Transformer with a higher number of parameters and with pre-trained weights of ImageNet 22-k it will cross the accuracy of the base paper. The results are shown in the table that include comparison with ResNet and the originally proposed Vision Transformer and it can be seen from the table that our proposed model performed competitively against different models which indicates

that the transformer can be a replacement for CNN models.

Method	mAP	Rank-1	Rank 5	Rank 10	Year
RPTM	87.4	96.2	98.5	-	2021
A Strong Baseline	87.1	97.0	-	-	2021
VehicleNet	83.41	96.78	-	-	2020
TransReID	78.6	95.9	-	-	2021
Cal	74.3	95.4	-	-	2021
Our	78.5	96.1	98.3	99.0	2022

Table 6.2: This table shows the comparison of the accuracy of our proposed model with various benchmark vehicle re-identification methods

Chapter 7

Conclusion

In this chapter, we will be discussing the concluding remarks and the contributions of this research. Moreover, the future direction based on the research conducted on vehicle re-identification will be discussed.

7.1 Concluding Remarks

Vehicle re-identification has become an important topic in the community of computer vision and a lot of research has been done in the past few years, but it is still a challenging task for the research community. As vehicle re-identification inherits some serious problems such as Limited publicly available datasets, viewpoint variance, various illumination conditions, occlusions, perspective/scale, quality of feed variations, difficult backgrounds, and shape changes. The biggest challenge of the vehicle re-identification is the difference within the class because of the different illumination and perspective changes and the similarity between the different classes e.g., vehicles by different manufacturers are very similar in appearance.

All of these can be managed efficiently with improvements in different areas. More information needed to be public or added to the dataset that will be useable for the public. For instance, VeRi-776 and VehicleID both are created by Chinese universities, and these are considered benchmark datasets for the community, but they do not provide video information that makes it hard to track vehicles in a wide

area. Besides, some datasets lack spatial-temporal information which limits the number of research being conducted based on spatial-temporal information. Most deep learning-based methods do not consider using of spatial-temporal information that can assist in real-time surveillance scenarios. spatial-temporal information also limits the number of computations being done by the system as the system only has to look for the vehicle in the next close camera, ignoring all other cameras this decreases the search time for the vehicle.

7.2 Future Work

The number of vehicles has increased in the past few years, it has led to an increase in the demand for surveillance, smart traffic management, and public safety. Many methods have been introduced to tackle this problem that was based on CNN, and CNN inherits some problems that a Vision Transformer can fix which have been proved in this research thesis. If vision transformers will widely be used in vehicle re-identification, then more efficient approaches based on transformers will be introduced that will be computationally inexpensive than CNNs and will have better representation learning power.

Besides, as datasets are a major issue in the vehicle-identification community, Vision Transformers can be applied on unsupervised baselines and that results in better accuracy.

References

[1] Gartner Says. 8.4 billion connected "things" will be in use in 2017, up 31 percent from 2016. 2017. *URL: https://www. gartner. com/newsroom/id/3598917*, 2018.

[2] Yi Zhouy and Ling Shao. Viewpoint-Aware Attentive Multi-view Inference for Vehicle Re-identification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00679.

[3] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2018. ISSN 15209210. doi: 10.1109/TMM.2017.2751966.

[4] Pedro Antonio Marin-Reyes, Luca Bergamini, Javier Lorenzo-Navarro, Andrea Palazzi, Simone Calderara, and Rita Cucchiara. Unsupervised vehicle re-identification using triplet networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018-June:166–171, 2018. ISSN 21607516. doi: 10.1109/CVPRW.2018.00030.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020. URL http://arxiv.org/abs/2010.11929.

[6] Ming Li, Xinming Huang, and Ziming Zhang. Self-supervised Geometric

Features Discovery via Interpretable Attention for Vehicle Re-Identification and Beyond. *Proceedings of the IEEE International Conference on Computer Vision*, pages 194–204, 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.00026.

[7] Hongbo Wang, Jiaying Hou, and Na Chen. A Survey of Vehicle Re-Identification Based on Deep Learning. *IEEE Access*, 7:172443–172469, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2956172.

[8] Tony Lindeberg. Scale Invariant Feature Transform. *Scholarpedia*, 7(5):10491, 2012. doi: 10.4249/scholarpedia.10491.

[9] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-identification. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:379–387, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.49.

[10] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized Near-duplicate Vehicle Re-identification. 1:3997–4005.

[11] Jinjia Peng, Huibing Wang, Tongtong Zhao, and Xianping Fu. Learning multi-region features for vehicle re-identification with context-based ranking method. *Neurocomputing*, 359:427–437, 2019. ISSN 18728286. doi: 10.1016/j.neucom.2019.06.013. URL https://doi.org/10.1016/j.neucom.2019.06.013.

[12] Xingan Ma, Kuan Zhu, Haiyun Guo, Jinqiao Wang, Min Huang, and Qinghai Miao. VEHICLE RE-IDENTIFICATION WITH REFINED PART MODEL University of Chinese Academy of Sciences , Beijing , China , 100049 National Laboratory of Pattern Recognition , Institute of Automation Chinese Academy of Sciences , Beijing , China , 100864. pages 1–4.

[13] Hongchao Li, Xianmin Lin, Aihua Zheng, Chenglong Li, Bin Luo, Ran He, and Amir Hussain. Attributes Guided Feature Learning for Vehicle Re-

Identification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021. ISSN 2471285X. doi: 10.1109/TETCI.2021.3127906.

[14] Peixiang Huang, Runhui Huang, Jianjie Huang, Rushi Yangchen, Zongyao He, Xiying Li, and Junzhou Chen. Deep feature fusion with multiple granularity for vehicle re-identification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2019-June, pages 80–88, 2019. ISBN 9781728125060. URL https://openaccess.thecvf.com/content_CVPRW_2019/papers/AICity/Huang_Deep_Feature_Fusion_with_Multiple_Granularity_for_Vehicle_Re-identification_CVPRW_2019_paper.pdf.

[15] Jin Hui Hou, Huan Qiang Zeng, Lei Cai, Jian Qing Zhu, and Jing Chen. Random occlusion assisted deep representation learning for vehicle re-identification. *Kongzhi Lilun Yu Yingyong/Control Theory and Applications*, 35(12):1725–1730, dec 2018. ISSN 10008152. doi: 10.7641/CTA.2018.80488.

[16] Saghir Ahmed Saghir Alfasly, Yongjian Hu, Tiancai Liang, Xiaofeng Jin, Qingli Zhao, and Beibei Liu. Variational Representation Learning for Vehicle Re-Identification. In *Proceedings - International Conference on Image Processing, ICIP*, volume 2019-Septe, pages 3118–3122, 2019. ISBN 9781538662496. doi: 10.1109/ICIP.2019.8803366. URL https://ieeexplore.ieee.org/abstract/document/8803366/.

[17] N Jiang, Y Xu, Z Zhou, W Wu 2018 25th IEEE International, and Undefined 2018. Multi-attribute driven vehicle re-identification with spatial-temporal re-ranking. *ieeexplore.ieee.org*. URL https://ieeexplore.ieee.org/abstract/document/8451776/.

[18] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, 2003. ISBN 0262025507. URL https://proceedings.neurips.cc/paper/2002/hash/c3e4035af2a1cde9f21e1ae1951ac80b-Abstract.html.

[19] Zakria, Jingye Cai, Jianhua Deng, Muhammad Umar Aftab, Muhammad Saddam Khokhar, and Rajesh Kumar. Efficient and deep vehicle re-identification using multi-level feature extraction. *Applied Sciences (Switzerland)*, 9(7), 2019. ISSN 20763417. doi: 10.3390/app9071291. URL https://www.mdpi.com/435486.

[20] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:2167–2175, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.238.

[21] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning Deep Neural Networks for Vehicle Re-ID with Visual-spatio-Temporal Path Proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 1918–1927, 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.210. URL http://openaccess.thecvf.com/content_iccv_2017/html/Shen_Learning_Deep_Neural_ICCV_2017_paper.html.

[22] Chao Cui, Nong Sang, Changxin Gao, and Lei Zou. Vehicle re-identification by fusing multiple deep neural networks. In *Proceedings of the 7th International Conference on Image Processing Theory, Tools and Applications, IPTA 2017*, volume 2018-Janua, pages 1–6, 2018. ISBN 9781538618417. doi: 10.1109/IPTA.2017.8310090. URL https://ieeexplore.ieee.org/abstract/document/8310090/.

[23] R Kuma, E Weill, F Aghdasi 2019 International Joint ..., and undefined 2019. Vehicle re-identification: an efficient baseline using triplet embedding. *ieeexplore.ieee.org*. URL https://ieeexplore.ieee.org/abstract/document/8852059/.

[24] Y Zhang, D Liu, ZJ Zha 2017 IEEE International Conference, and undefined 2017. Improving triplet-wise training of convolutional neural network for ve-

hicle re-identification. *ieeexplore.ieee.org*. URL https://ieeexplore.ieee.org/abstract/document/8019491/.

[25] Y Li, Y Li, H Yan, J Liu 2017 IEEE International Conference On, and undefined 2017. Deep joint discriminative learning for vehicle re-identification and retrieval. *ieeexplore.ieee.org*. URL https://ieeexplore.ieee.org/abstract/document/8296310/.

[26] R. M.S. Bashir, M. Shahzad, and M. M. Fraz. VR-PROUD: Vehicle Re-identification using PROgressive Unsupervised Deep architecture. *Pattern Recognition*, 90:52–65, 2019. ISSN 00313203. doi: 10.1016/j.patcog.2019.01.008. URL https://doi.org/10.1016/j.patcog.2019.01.008.

[27] Raja Muhammad Saad Bashir, Muhammad Shahzad, and Muhammad Moazam 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)*, volume 11241 LNCS, pages 286–295. Springer Verlag, 2018. ISBN 9783030038007. doi: 10.1007/978-3-030-03801-4_26.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. ISSN 15577317. doi: 10.1145/3422622.

[29] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based Object Re-Identification. *Proceedings of the IEEE International Conference on Computer Vision*, pages 14993–15002, 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.01474.

[30] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A Strong Baseline and Batch Normalization Neck for Deep Person Re-Identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, oct 2020. ISSN 19410077. doi: 10.1109/TMM.2019.2958756.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9992–10002, 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.00986.

Appendix A

First Appendix

A.1 Accuracy Graph of Vehcile Re-Identification on VeRi

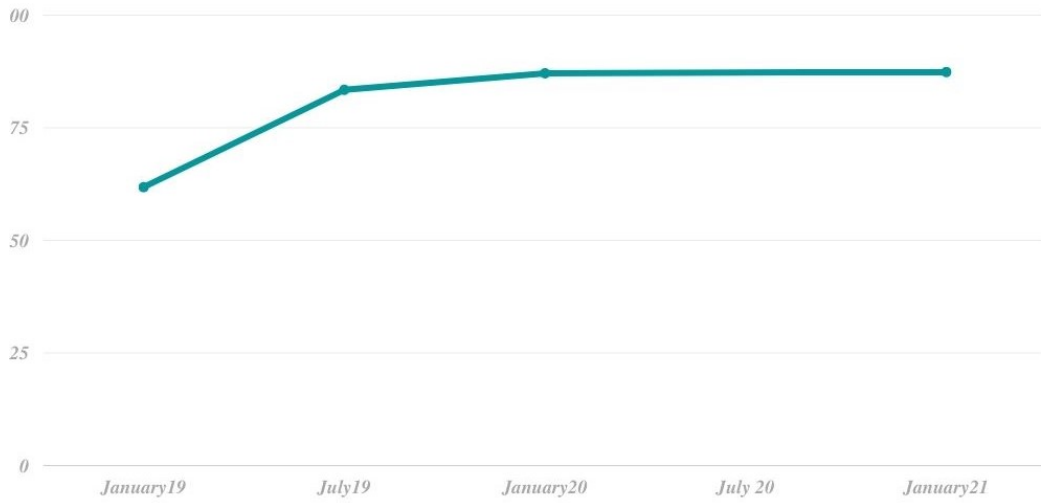

Figure A.1: This graph is indicating the accuracy of the state-of-the-art models on VeRi-776 dataset over the years.