

Analysis of Failures in State-of-the-Art Person ReID Models

by

Kristian Suhartono

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in Industrial Engineering and Decision Analytics

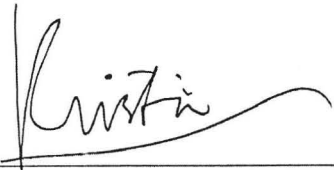
November 2021, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

A handwritten signature in black ink, appearing to read 'Kristian', is written over a horizontal line.

Kristian Suhartono
29 November 2021

Analysis of Failures in State-of-the-Art Person ReID Models

by

Kristian Suhartono

This is to certify that I have examined the above MPhil thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.



Prof. Richard Hau Yue So (Supervisor)



Prof. Guillermo Gallego (Head of the Department)

Department of Industrial Engineering and Decision Analytics

29 November 2021

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor Professor Richard H. Y. So who has guided me through the journey that is my master study. He has been very supportive, encouraging, and challenging me throughout our time together, pushing me to always push on and question the scientific boundaries of the work that we do. Without him, I would not have been able to create this thesis.

Secondly, I would like to thank all my friends in HKUST who supported me as I prepared this thesis. Especially Dr. Nick Chin, Teric Chan, Eddie Lau, Kyle Wong, Men Yixin, and Phoebe Ching for all their help in the project that created the inspiration for this thesis. I will always be grateful for all your help and advice throughout the time that I spent with all of you, every one of you inspires me. I would also like to thank my research group members for all their support and companionship throughout my time with them. All of you have given me amazing and memorable experiences that I will cherish for the rest of my life.

Thirdly, I would like to thank the people in my church community that have walked with me through the process of my studies. All of you have continuously encouraged me and challenged me to grow throughout this journey. Without all of you who kept on reminding me of the truth and encouraging me, I would not be where I am today and this thesis would not be complete without all of your help. Special thanks to Pastor Bo Zhu for mentoring me and guiding me as a person and challenging me to always grow and not be complacent.

Finally, I would like to thank my family for caring for me and supporting me throughout the challenging years in Hong Kong. To my mother, Monica Jannie Andajani, and my father, Kurniawan Suhartono, for all the lessons that you continue to impart into my life. Thank you for being patient with me and continuously supporting me this whole time.

Table of Contents

Title Page	i
Authorization Page	ii
Signature Page	iii
Acknowledgments	iv
Table of Contents	v
List of Figures	vii
List of Tables	ix
Abstract	x
Chapter 1: Introduction	1
Summary	1
1.1 Background and Motivation	1
1.2 Contributions	2
1.3 Thesis Structure	3
Chapter 2: Literature Review	4
Summary	4
2.1 Person Re-identification (ReID)	4
2.1.1 Online Person ReID	5
2.1.2 Open-World Person ReID	6
2.2 Occluded Person ReID	7
2.3 Research Gaps	8
Chapter 3: Implementation and Optimization of an Online Person ReID	
System in a Screening System	9
Summary	9
3.1 Introduction	9
3.2 Methods	10
3.2.1 Equipment and System Design	10
3.2.2 Image Preprocessing	12
3.2.3 CNN Module	13
3.2.4 Post Processing	13
3.2.5 Metrics	14
3.3 Results and Analysis	16

3.4	Discussions and Conclusions	19
Chapter 4: Understanding the Cause of Failure in State-of-the-Art Person ReID Networks 20		
	Summary	20
4.1	Introduction	20
4.2	Methods	20
4.2.1	Initial Benchmark	21
4.2.2	Experiment 1: Examining Model Failure Cases	21
4.2.3	Experiment 2: Exploring the Impact of Occlusion	22
4.2.4	Experiment 3: Exploring Whole and Occluded Images and Their Impact on Model Performance	23
4.3	Results and Analysis	24
4.3.1	Benchmark Results	24
4.3.2	Experiment 1 Results	26
4.3.3	Experiment 2 Results	32
4.3.4	Experiment 3 Results	33
4.4	Discussions and Conclusions	35
Chapter 5: Impact of Occluded Training Samples in Occlusion Situations 37		
	Summary	37
5.1	Introduction	37
5.2	Methods	37
5.2.1	Dataset and Models	37
5.2.2	Experiment Setup	38
5.3	Results and Analysis	38
5.4	Discussions and Conclusions	40
Chapter 6: Conclusions, Limitations and Future Works 42		
6.1	Conclusions	42
6.2	Limitations and Future Works	42
References 44		
Chapter A: Comparison of Models 47		

List of Figures

3.1	Camera setup of the system which shows 6 different viewing angles over the whole site. The varying angles and viewpoints are designed to allow full coverage of the area where people walk through.	11
3.2	Result of testing our base models on our manually labeled personal dataset. CMC-1 accuracy, which measures the rate at which the model's best gallery match's identity is the same as the query. Case 1-3 are easy cases where subjects are mostly unobstructed and do not cross each other's paths. Case 4-6 are hard cases where subjects often cross paths with one another, causing occlusions to happen.	17
3.3	Sample images from our manually labeled personal dataset. Top row: case 1-3, Bottom row: case 4-6.	18
4.1	Effect of varying inference batch size on model mAP. mAP is a metric of ReID model performance that measures its precision in fetching images of the same identity	24
4.2	mAP and speed of inference for each model with a batch size of 128. mAP is a metric of model performance	25
4.3	Size of intersection sets of ReID failure cases between the models with the CMC-1 metric. Intersection sets contain the images where all the models listed are unable to meet the CMC criteria. CMC-n checks for whether the correct identity match is within the top n matches. Size of each model's failure set is shown on the left bar.	26
4.4	Size of intersection sets of ReID failure cases between the models with the CMC-5 metric. Intersection sets contain the images where all the models listed are unable to meet the CMC criteria. CMC-n checks for whether the correct identity match is within the top n matches. Size of each model's failure set is shown on the left bar.	27
4.5	Size of intersection sets of ReID failure cases between the models with the CMC-10 metric. Intersection sets contain the images where all the models listed are unable to meet the CMC criteria. CMC-n checks for whether the correct identity match is within the top n matches. Size of each model's failure set is shown on the left bar.	27
4.6	Size of intersection sets of ReID failure cases between the models with the CMC-1 metric when same camera matches are allowed. Size of each model's failure set is shown on the left bar.	28

4.7	Size of intersection sets of ReID failure cases between the models with the CMC-5 metric when same camera matches are allowed. Size of each model's failure set is shown on the left bar.	28
4.8	Size of intersection sets of ReID failure cases between the models with the CMC-10 metric when same camera matches are allowed. Size of each model's failure set is shown on the left bar.	29
4.9	The query images in the CMC-1 failure cases intersection set of the 4 selected models.	30
4.10	Sample results from the Feature Visualizer on the regions that the 4 selected models pay attention to when creating the image's feature representation .	31
4.11	Result of testing SOTA models and ResNet-18 as a benchmark on our manually labeled personal dataset. CMC-1 accuracy measures the rate at which the model's best gallery match's identity is the same as the query. Case 1-3 are easy cases where subjects are mostly unobstructed and do not cross each other's paths. Case 4-6 are hard cases where subjects often cross paths with one another, causing occlusions to happen.	32
4.12	Result of testing SOTA models and ResNet-18 as a benchmark on the OccludedReid dataset. The contents of the query and gallery set are varied according to the settings in the Y axis to measure the effect of occlusion. CMC-1 accuracy measures the rate at which the model's best gallery match's identity is the same as the query.	33
4.13	Effect of changing the ratio of occluded and whole images in the gallery set on model's performance on OccludedReID Dataset with different query types	34
5.1	Comparing the effects of having occluded training data to the model's CMC-1 accuracy on the OccludedReid test set. CMC-1 is a metric for a ReID model's performance.	39
5.2	Comparing the effects of having occluded training data to the model's CMC-1 accuracy on the Personal dataset. CMC-1 is a metric for a ReID model's performance.	40

List of Tables

3.1	Performance of models on the test set of Market-1501 dataset after the training process, mAP (mean average precision) is the measured likelihood of all images of a person of the same identity being the top matches for the query, CMC-n measures the accuracy of a gallery match with the correct identity being in the top n matches	17
3.2	Performance of the system on simulated crowd behavior cases under different test conditions. Speed denotes the walking speed of the subjects in the test case, N/A cases were not simulated.	18
4.1	Effect of adding extra images to the gallery set on model performance with whole query images, accuracy drops when there are no whole image samples in the gallery. The numbers in the table are the percentage CMC-1 accuracies of the models. CMC-1 accuracy denotes the model's rate of having the best gallery match as the true identity match with the query. .	35
4.2	Effect of adding extra images to the gallery set on model performance with occluded query images. The numbers in the table are the percentage CMC-1 accuracies of the models. CMC-1 accuracy denotes the model's rate of having the best gallery match as the true identity match with the query. Note the low accuracy when compared to the results in Table 4.1	35
5.1	Performance of selected models on the test set of OccludedDuke dataset after the training process, the mAP and CMC-1 are metrics for model performance.	38

Analysis of Failures in State-of-the-Art Person ReID Models

by Kristian Suhartono

Department of Industrial Engineering and Decision Analytics
The Hong Kong University of Science and Technology

Abstract

Humans can reliably find a person through a crowd but computer programs often fail. Even with the help of deep learning, person re-identification (ReID) networks can fail. Studies on reasons for failure have been few. This is because these networks have high dimensional complexity [1]. The lack of understanding limits our ability to improve the ReID networks. This study developed and implemented a real-time person ReID system. The systems were tested to determine the boundary conditions between success and failure. Paths to failure in state-of-the-art deep learning ReID models were analyzed. Findings open up the possibility of future improvements.

We implemented and optimized a real-time ReID system as part of larger screening systems. The systems were tested and deployed at border control points, specifically the Hong Kong International Airport. Test results indicated a discrepancy between the measured accuracy of a model on the training data and on-site performance in real settings. The issue identified was occlusion.

In parallel, we explored how recent state-of-the-art ReID networks decompose and reconstruct the image information and followed the design-of-experiment technique to study and examine the network mechanisms associated with ReID failures. Convergingly, we discovered occlusion also plays a significant part in the failure of the models. Surprisingly, using an occluded query image to search for an occluded match did not improve the performance. Ensuring the query image is not occluded greatly improved model accuracy.

Furthermore, we discovered that retraining using training data that contained occluded samples improved the model accuracy for occluded images but degraded its performance for whole (unoccluded) images. Possible applications of the findings are future enhancements to ReID networks through improvement on the training dataset or through different network architecture designs

Chapter 1

Introduction

Summary

In this chapter, the background and motivation of the study and the research contributions are briefly introduced. The contributions of the thesis are outlined in the next section to summarize the major pillars in this thesis. At the end of the chapter, the structure of the thesis is presented to describe the contents of each chapter in the thesis.

1.1 Background and Motivation

Camera systems have been widely used as a way to monitor a certain area, and this creates a demand for a more complex camera system that is able to capture information from different angles and positions. Oftentimes such systems still require human supervision that watches the output of the cameras and extracts any information that they need to make a decision. This is often very laborious and prone to error as the system becomes more complex it makes it harder and harder for humans to monitor all the different viewpoints.

As artificial intelligence (AI) technologies expand, the field of computer vision developed multiple techniques to assist with the need for automated systems that are able to extract information from multiple sources. One of the main fields of research focuses on Person Re-identification (ReID) which entails the problem of matching a person's identities across multiple camera viewpoints. However, even though research in the field has been growing in recent years there hasn't been as many applications of the technology in real-life situations. There are several key challenges that make this adoption in real situations difficult. Firstly, the need for online responses from the system creates a lot of constraints for the AI method. Approaches in Person ReID have been designed to be run offline where time is not a major concern, thus increasingly complex models have been developed in order to achieve better accuracy rates. However, in a real situation, such systems usually need to give real-time responses that will help humans make a decision. Secondly, most of the approaches require extra data in order to adapt to a certain real life implementation. Information such as site-specific angles, lighting, background noise are things that models often need to learn to filter out to increase their performance in a certain situation. However this kind of data is often difficult to obtain, thus severely impacting the performance of these AI approaches.

This thesis implemented a Person ReID system that takes into account these challenges and optimized different parts of the system so that it is able to achieve considerable

performance in an online setting. Operating under the assumption that the models won't have extra training data, we developed other methods to improve the capabilities of the system in matching the identities of people across different camera viewpoints. However, during the testing process, we discovered a discrepancy between the performance of the model on the training data as compared to its performance in a real-life situation. There were two problems that we identified that we believed could have caused the difference in performance. Firstly, the way the model extracts information from input data is often a black-box approach that is not easy to understand at first glance. Most AI systems, especially deep learning approaches, become increasingly complex which makes it difficult for humans to interpret the model's decision-making process. Secondly, the variance between training data and real situation data causes AI models to be unable to arrive at a generalizable solution that works in all situations. Thus there needs to be an optimization on the training data that better represents the real situation so that the model is able to find the solution that works in such situations.

To summarize, in this thesis, we implemented a system that identifies a person as they move across an area of interest that is monitored by different cameras. Furthermore, we optimized that system to achieve a more reasonable degree of accuracy and reliability using data processing methods. Secondly, we explored the reason behind the failures of the models, especially state-of-the-art approaches to understanding what is the main challenge that is keeping these models from being able to be used in a real situation. Finally, we investigated the possibility of enhancing such models by changing their training data to more accurately represent real-life situations.

1.2 Contributions

In this thesis, we developed and tested a Person ReID pipeline that is part of a temperature screening system. Due to the challenges of the online and open-world environment, we also optimized the system to work under the constraints of the environment. Furthermore, we explored the reason behind the failure cases of state-of-the-art approaches. Finally, we studied the effects of adding occluded samples to the training data of a model and how it performs on different evaluation datasets. The main contributions of the thesis are as follows:

1. **Implementation and optimization of a person ReID pipeline in an online and open-world setting:** We implemented a ReID system that was tested on simulated cases of crowd behavior. Deep learning approaches and algorithms were utilized to develop the system, resulting in a 60% success rate in our test cases.
2. **Failure causes in state-of-the-art person ReID approaches:**
 - (a) Occlusion has a significant impact on model performance, in occluded cases models are only able to achieve 30-55% CMC-1 accuracy. CMC-1 accuracy

is a metric that is used to measure how well a model can correctly match a person’s identity by checking whether the best match between the query and the gallery are of the same identity.

- (b) Having a whole body query image can cause on average a 20.7% accuracy increase in the CMC-1 accuracy of the model, even when the gallery images are occluded.
- (c) Having a small sample of whole-body images for a certain identity in the gallery increases the CMC-1 accuracy by 29.9% on average when the query image is whole.
- (d) A set of experiments for benchmarking model performance in handling occlusion cases

3. **Effects of modifying training data on the performance of state-of-the-art person ReID approaches:** Occluded samples in the training data improve part-focused models’ capabilities in handling occlusion cases on average by 6.72% and 13.04% on our 2 evaluation datasets.

1.3 Thesis Structure

The thesis is structured into several chapters, the contents of each chapter are as follows:

Chapter 1 provides an introduction to the background and motivation of the thesis and gives a summary of the thesis’s contributions and structure.

Chapter 2 summarizes different information about the field of Person ReID as a whole, particularly on state-of-the-art approaches to Person ReID and open-world, online, and occluded Person ReID.

Chapter 3 details the design of an Online Person ReID pipeline that is a part of a person screening system and the optimizations that were applied to enhance the system performance.

Chapter 4 studies the failure cases in a Person ReID system, particularly in state-of-the-art approaches.

Chapter 5 explores the strategy of modifying the training data to enhance the performance of a Person ReID system.

Chapter 6 concludes the findings of this thesis and suggests possible directions for future works.

Chapter 2

Literature Review

Summary

This chapter presents a critical review of the field of person ReID, particularly its more state-of-the-art approaches that are performing well on benchmark datasets. Furthermore, it reviews aspects of the field of person ReID, particularly online, open-world, and occluded person ReID which are challenges of a real-life person ReID system. In the end, the chapter presents a summary of research gaps.

2.1 Person Re-identification (ReID)

Person Re-identification is often defined as the task of retrieving identity matches of a certain person across images from multiple camera viewpoints. Typically, there is an image of a target person that we are looking for (query) and a series of images that may or may not contain a person with the same identity (gallery). A detailed history of the field can be found in [2]. In recent years deep learning approaches have often been employed to solve this problem [1] [3] [4]. Such approaches have performed well as measured on benchmark datasets such as [5], [6], and [7]. However, there are still major challenges that the field is still working on solving. Particularly low image resolutions, lighting differences, variations between human poses, occlusions, and heterogeneous modalities [8]. This chapter presents a critical review of the field of person ReID, particularly its more state-of-the-art approaches that are performing well on benchmark datasets. Furthermore, it reviews aspects of the field of person ReID, particularly online, open-world, and occluded person ReID which are challenges of a real-life person ReID system. In the end, the chapter presents a summary of research gaps.

Several deep learning based approaches have attempted to tackle those challenges by developing different methods to extract information from a picture that can be separated into 3 categories. The first one is to extract global features from the image, taking the image as a whole and letting the model recognize distinctive patterns that help differentiate identities. This approach is used by [9] and [10], the resulting models often take very general patterns that allow them to be more flexible to noise. However, this also limits the accuracy of the model as it is less capable in picking up more detailed or minute features such as shoe patterns that can be significant in differentiating identities. The second category is to focus on local features using targeted attention mechanisms. This can be seen in [11][12][13][14][15] where they either focus solely on local features or combine the information from global and local features to differentiate identities. Models like this are

more accurate than the ones that just focus on global features, yet they are more vulnerable to noise. Domain variance between the training data and test data often causes the models to have a considerable drop in accuracy. Occlusion of certain body parts and incomplete images are something that becomes hard to handle as when body parts are missing, it causes an unfair comparison between images. The third category focuses on adding auxiliary features to the model’s decision making process. That is to modify the information that is being given to the model through data enrichment or manipulation and also approaches that add extra data dimensions. [16] and [17] are examples of such models and it is seen that they are less prone to noise and yet have a comparable performance to the part based approaches. This however often comes at the cost of making the models more complex or needing more training data which can cause it to become unsuitable for real time use.

As a whole, these approaches are often trained and tested on a benchmark dataset, however these datasets often represent a closed set problem for ReID that means most of the possible cases are represented in it’s training data. Whereas in real life situations, there is usually different challenges to be tackled due to the requirement of real time processing and the possibility of data that wasn’t in the training dataset to appear in the use case. These are often represented as online and open-world person ReID.

2.1.1 Online Person ReID

In real situations, ReID is often used as a core component of a real time system that runs with a human observer looking at the output of the system. This creates the need for ReID approaches to not be too costly to allow human observers to be able to directly react to the information that is presented to them by the ReID system. The term online person ReID defines this setting, where a ReID system will need to complete its task within a certain time window. Thus creating time and model complexity constraints for the approaches of models used in this setting.

There haven’t been too many unique approaches on how the problem of online person ReID is being tackled, the main technique that is used is reducing the complexity of the person ReID models. [18][19][20] substitutes the regular real value representation of the output of feature extraction into binary codes. Essentially reducing the complexity of computing the feature representation which also allows for faster and more efficient algorithms when comparing the distance between sets of features. This reduces the amount of time taken to compare the extracted features while maintaining similar accuracy to other state of the art approaches. Another approach in [21] attempts to change up the ReID backbone to reduce the number of parameters and also power consumption since the models are designed to be used in edge devices. This showed that it is possible to use less complex models as the backbone for the ReID network and still maintain a high degree of accuracy.

However as compared to the number of research that employs state-of-the-art and

complex models, online ReID hasn't received as much attention. Most of the approaches detailed that are state-of-the-art have increased model complexity, using deeper layers or more sources of information to increase the available information that can be used to make a decision. Whereas most of the online person ReID solutions boil down such approaches to the bare minimum to ensure that only the most significant features are being used to optimize the model's efficiency. This suggests that the direction of the state-of-the-art is the opposite of most online person ReID solutions. Furthermore, most approaches to person ReID rely on ResNet as the backbone of the model. This results in the inference time that is needed for the model to process images to not change too much. This is certainly due to the great success of ResNet based CNNs [1] in extracting features from images for contexts such as person ReID. However, this creates the constraint of having to use these networks again and again which limits the speed or throughput of a ReID system.

2.1.2 Open-World Person ReID

So far, most of the approaches that have been described are focused on the problem of closed-world person ReID. This means that there are several assumptions about the setting of the ReID system as is defined in [8]. The main base assumption that differentiates closed-world ReID and open-world ReID is the existence of the target person in the gallery set [22]. In open-world ReID, there are occurrences where the target identity doesn't appear in the search space. Most of the closed-world approaches usually return a ranked list of similarities between the query and gallery image, which creates a problem in closed-world ReID because the best match (rank-1) may not even be remotely similar to the query image. Such a case often occurs in real life situations especially when the query person disappears from all camera viewpoints. Another challenging situation arises to the fact that in an open-world ReID setting, the appearance of a person may change especially if the system is being used to monitor people in a large area or a long period of time. This essentially trivializes most person ReID approaches as they rely heavily on the appearance of a person to obtain information about the identity of the person.

A survey of open-world ReID approaches [22] summarized that open-world person ReID is challenging and there aren't a lot of existing metrics and data that allow researchers to directly compare their approaches to each other. Whereas in the closed-world setting where there are many datasets, baselines, and evaluation metrics. This creates the lack of emphasis on this aspect of person ReID even though this setting is the one that more closely resembles real life situations. The approaches that are reviewed in the survey are not as recent as other state-of-the-art approaches and also the setting of those studies aren't directly defined as "open-world" ReID although they share similarities with characteristics of open-world ReID. This indicates that this aspect of ReID is still young and is still far from being solved.

2.2 Occluded Person ReID

When using camera systems to monitor a certain area, the number of people that are in the area is often unpredictable. When the number of people on the scene is sparse, that means the cameras are usually able to get whole body pictures of a certain person easily. But when there are dense or large crowds, this usually causes different people to appear occluded in the images that are taken. This situation is often exacerbated by other obstacles that are part of the scene, i.e. furniture, decoration, crowd control measures. This situation where there is a high possibility of the query images or gallery images to be corrupted by some sort of obstruction encapsulates the setting of Occluded Person ReID.

The main challenge with occluded person ReID lies in these occlusions causing missing information that are often key to differentiating identities. This causes unfair comparisons where the model isn't able to extract information from all the key places that it can usually extract from when it has the whole image. One of the first approaches that addressed this is [23] where they looked at part based information to counteract the fact that global information may not be consistent and then add complementary global information to give spatial cues so that relevant local features can be compared. This approach is very computationally costly as the features needed to construct the person's identity needs to be repeatedly computed based on which parts are available. [24] remedies this by creating an alignment free sparse feature representation thus removing the need to repeatedly recompute the person representation of the query. They further improve this approach by using foreground and background awareness to focus more on key features and make the model less susceptible to background noise in [25]. However, these approaches often take up the assumption that the gallery images are whole, which isn't guaranteed in a lot of occlusion situations. [26] appropriately models this situation and tries to solve it through restricting the comparison process to only use the available body parts that can be compared. [27] takes this another step further by removing the limitation that a certain body part must be compared to another body part and instead letting the model learn a higher order information that can merge different parts together and compare the features at that level. To summarize, as a whole the field has been focusing on how to create a fairer representation of the person given the missing information. However, these methods imply that extra information is needed which increases the model complexity and thus slows down the models. Even then, with cases where there are heavier occlusions, it seems that these models still struggle to correctly match the correct identity. [27] reported a 55.1% accuracy in matching identities between the query and gallery images.

There are several datasets that have been created to model this situation, OccludedReID [28] and PartialReID [23] are evaluation datasets that contain 5 whole images and 5 occluded images of 200 and 60 identities respectively. Occluded-DukeMTMC [26] is a re-split of the original DukeMTMC dataset [6] such that the queries of its evaluation set are all occluded images and the gallery consists of at least 10% occluded images. However, the sizes of these occluded datasets are very small, the most well known benchmark for

Person ReID is Market1501 which contains 32,668 images from 1501 identities. Furthermore, to the best of our knowledge, there exists no dataset that only contains occluded images that have been used for person ReID. This discrepancy in dataset size highlights how Occluded Person ReID needs the inavailability of data sources that are able to model the problem.

2.3 Research Gaps

From the literature review, we concluded that a large portion of ReID research has been mainly focused on making models more complex and better in the closed-world ReID problem. However, this setting doesn't represent real situations that often have the characteristics of being an online, open-world, and occluded person ReID. Each of these areas of ReID haven't gotten a lot of attention in recent years of research. The research gaps that we identified from the literature are summarized as follows:

- There are few studies regarding which have ReID systems that are deployed and tested in an online and open-world environment.
- There are no studies that explore the effect of occlusion on model performance, in particular when compared to the whole image setting.
- There are few studies on the effects of adding occluded data into the training dataset of a model.

Chapter 3

Implementation and Optimization of an Online Person ReID System in a Screening System

Summary

This chapter details the implementation of a person ReID system as a part of a person screening system that has been tested under simulated crowd behavior cases. It also details the challenges with such a system and the steps that were taken to optimize the system.

3.1 Introduction

Monitoring incoming information from a large number of cameras is not an easy task and often becomes laborious and is prone to human error. The larger the number of cameras, the more difficult the situation becomes for the human operator to be able to keep track of all the different angles and viewpoints, especially when the field-of-views don't overlap. Especially when there is a requirement of identifying and tracking someone as they walk through the different camera viewpoints. Such is the case with our setup at the Hong Kong International Airport[29], where checkpoints are set up where operators from the Department of Health will perform checks on traveler's temperatures. In such a system, multiple cameras are used by the operators to identify a certain person that has fever symptoms which is often an indicator of disease. The identified person is then tracked as they walk through the checkpoint area and is then stopped by the operator for further examination. To cope with this need, human operators have been watching screens and doing eyeball checks to identify and track a certain person as they move through the camera views. This limits them in terms of the number of people that they can track, and also that the process requires multiple people to be involved as the operator that tracks cannot leave the monitoring station, resulting in needing other operators to catch and stop identified suspects. Furthermore, suspects can disappear due to occlusion especially when there are larger crowds, and also when using nearby facilities. Given this situation, there is a great need to develop a system that can reduce and even completely replace the human factor using deep learning approaches that have proven to be successful in dealing with crowd monitoring.

We propose a Person ReID system that will identify targeted people as they walk through an area over multiple camera viewpoints. The ReID system is part of a fever

screening system where the ReID system is mainly used to reidentify people as they appear in multiple camera viewpoints, and also in maintaining their identity as a suspect even if they disappear due to occlusion or because they left the scene. The ReID system consists of an image preprocessor, a convolutional neural network (CNN) for identity matching, and a post-processing module.

3.2 Methods

In this section, we will describe the equipment that we use that supplies input data to the system. We also describe the process through which we process that information to output the identity of each person in all the camera views. We also detail how we measure the success of our system.

3.2.1 Equipment and System Design

The ReID system is a module that is part of the screening system that is used to perform fever screening on people that pass by a certain area. The system runs on two GPU servers where each GPU server is outfitted with 2x Intel® Xeon® Gold 6136 Processor (3.00 GHz) and 6 x NVIDIA (INNO3D) GeForce RTX 2080Ti Jet, and 256GB RAM. Each server processes data that originates from 3 color cameras that consist of 2 units of 8MP Hikvision Smart Network Box Camera and 1 unit of 12MP Hikvision Smart Network Box Camera. The cameras are arranged to capture different angles of sections of the monitored area to maximize camera coverage, especially when people disappear under occlusion. All the cameras face the direction where people come from to get a forward-facing view of the incoming people as this is required to capture their thermal data. This setup ensures that the images that are fed to the ReID system to be mostly forward-facing images. The images are captured frame by frame from the video stream of each camera.



Figure 3.1: Camera setup of the system which shows 6 different viewing angles over the whole site. The varying angles and viewpoints are designed to allow full coverage of the area where people walk through.

The screening system performs several tasks, but the work on this thesis focuses on the ReID system. The other modules are person detection, fever detection, and tracking. All the other modules process the data before the ReID module, which performs the final merging of information and outputs a list of images of people with identity-specific information. This entire process is required to run at approximately 10FPS due to the real-time system. As such the system is designed to be a pipeline where each module can occupy the given duration of 100ms before needing to complete the task and pass the information to the next module.

For the ReID system itself, all the data processing is done on one server which we call the master server. Data is shared between the servers through the use of a shared LAN network. In this server, one of the NVIDIA RTX 2080Ti GPU is dedicated solely for the use of the ReID system while the rest are used by other functions of the overall screening system. The CPU is being shared by the ReID system with all the other processes that are a part of the system through parallel processing.

The overall flow for the ReID system is as follows:

1. Input data from the 6 cameras are given to the ReID process.
2. Image preprocessing techniques are run on all the images.
3. Images are batched and sent to the GPU and the ReID model runs inference to extract identity features.
4. Extracted features are compared to match identities based on images.
5. Post-processing algorithms are applied to the results.
6. Output with identity numbers matched to corresponding images is returned.

3.2.2 Image Preprocessing

Images captured by the camera system and passed into the ReID module have two problems. The first one is that they come in various resolutions. The images are taken by taking the pixel RGB data inside a minimal bounding box patch that is drawn around the person based on the person's key points. The absolute size of the image patch differs according to the proximity of the person to the camera, the closer the person is, the larger the dimension of the image. However, due to the implementation of CNNs, images need to be resized to specific dimensions that can fit the convolutional layers. As such, we address this by resizing all given images to 256x128 pixels (height x width). The second problem is the camera position and environment situation that create a variety of lighting conditions. As can be seen in Figure 3.1, the lighting of the scene is not uniform, resulting in changes in how color appears in the images. As such, we first need to normalize the images that are passed into the system. We do this in 2 steps, first we apply the following algorithm to each image that is passed to the system:

1. Apply the OpenCV Histogram Equalization [30] to the image.
2. Get the mean pixel value of the image
3. Get the difference between the mean pixel value to a target mean pixel value c
4. Add the value c to all pixels in the image, value capped between 0-255

The value c is manually tuned using search methods on a predefined set of images that are picked out from actual test cases. Secondly, as per the standard defined by PyTorch, images that are used as an input for CNN models that are trained on the pre-trained models expect the images to be normalized according to the ImageNet mean and standard deviation [31]. As such we apply this to all the images before passing the data to the models.

3.2.3 CNN Module

The CNN module is divided into 2 parts, the data merging unit, and the CNN model. The data merging unit performs a merge on the identities based on the position of the person on the scene. As part of each person’s information, the merging unit receives the x-y position of the person on the floor map. A merging process is conducted when the closest distance between two people is below a certain threshold t which is tuned qualitatively by observing the merging result on our data. Using t , we minimize the number of false-positive merges that are caused by incorrectly picking out people that are close by in different cameras. This merging process is done to minimize the number of images that exist as the CNN model has a maximum limit on the number of images it can run inference if we want to keep the inference time under 100ms.

The CNN model is built using the architecture defined by [9] that uses a ResNet-50 network [1] as the backbone network that performs feature extraction. The model extracts features from the image and represents the image as a 1D vector of real values that explain the major descriptors of the identity of the person in the image. The vectors of the image that is defined as the query (the suspect that we are looking for) are then compared to the vectors of the image in the gallery (all the other person images from the frame). The vectors are compared using cosine distance to measure the distance between the two vectors, and the distance value is then used to rank the difference between vector pairs. Because we use cosine distance, the value is scaled between 0 to 1 where 1 signifies a perfect match between the 2 vectors. The vector pair with the highest distance is taken as the best ReID match which is chosen as the matching identity from the gallery set. The model is trained on the Market-1501 dataset on top of the base pretrained weights provided by PyTorch and then directly used in the system without any extra training data. The training process follows the process used by [9] which is commonly used to train all ReID networks.

We performed tests where we replaced the ResNet-50 with Resnet-18, AlexNet [32], and MobileNet [33] and the result is detailed in Section 3.3. Due to the requirements of the system, we had to optimize to ensure that the system is able to run the inference process under 100 ms. To accomplish this, we selected the model with the ResNet-18 backbone that has a slightly worse performance but is 1.9x faster than the ResNet-50. This is mainly done to increase the number of images that we can run inference on to allow the system to support ReID on a larger number of people.

3.2.4 Post Processing

The result of the CNN module that performs ReID is not always correct as can be seen in Section 3.3 but to ensure the robustness of the system and make sure that human operators aren’t confused by the identities constantly changing, we apply adaptive thresholding and an identity mapping system. This is done to minimize the impact of noisy matches due

to false ReID matches.

Adaptive Thresholding

We apply a hard threshold that is qualitatively optimized to ensure that only match results where the model is confident that the query and the best gallery match is similar are stored in the memory system. We also applied adaptive thresholding that looks at the entire gallery and checks the gap between the 1st place match and the 2nd place match. We observed that when the gap is large, this means that the model is confident that the query image is very similar to the 1st match and very different from the 2nd match. As such, we apply the following algorithm:

Algorithm 1 Adaptive Thresholding

```
if  $gap \geq gapthresh$  then  
     $threshold \leftarrow threshold + modifier$   
end if
```

Here we qualitatively optimize the *gapthresh* and *modifier* values to ensure that when the gap between matches is large, we relax the threshold to allow lower-quality matches through the threshold. We observed that this resulted in more true positive matches while minimizing the number of false positives.

Identity Mapping System

The identity mapping system is built to tackle the problem of the targets not existing in the gallery set due to occlusion or the subject completely leaving the field of view of the camera. This creates a memory system to ensure that when the same person reappears, the old suspect person’s identity will be matched to that person instead of creating a new identity for the person that reappeared. The mapping system also minimizes the effect of false matches by building a sort of memory when a repeated id is matched to a certain number. This is all done to ensure that to the human operator, the same person’s identity would always reappear to ensure consistent identification. The algorithm is as follows:

3.2.5 Metrics

The models’ performance is measured using metrics that are mainly used to measure the performance of ReID models. They are the mean Average Precision (mAP) that measures the ability of the model in making sure that all the top matches of the model are precise of the same identity. For example, the gallery set has 4 images that are labeled as the same identity as the query set, if the top 4 matches from the whole gallery set are those 4 images, then for this case it is given an average precision (AP) of 1. The AP value of all the queries is calculated and the mean value is taken as the mAP value. CMC-n is

Algorithm 2 Identity Mapping System

```
if person matched then
  if  $memory[id_{query}][id_{gallery}]$  then
     $memory[id_{query}][id_{gallery}] \leftarrow memory[id_{query}][id_{gallery}] + 1$ 
    if  $memory[id_{query}][id_{gallery}] \geq 20$  then
      reset  $memory[id_{query}]$ 
       $memory[id_{query}][id_{gallery}] \leftarrow 5$ 
    end if
  else
     $memory[id_{query}][id_{gallery}] \leftarrow 1$ 
  end if
end if
```

used to measure the general ability of the model by checking whether the true match is in the top n matches. If it is, then a value of 1 is assigned, otherwise, it is given the value of 0. The final CMC- n value is obtained by summing up all the CMC- n values of all the queries and then averaging it.

The system's success is measured using a set of 9 simulated crowd behavior cases using 6 test subjects (if the case specifies the number of people, then the number is used). Of the 6 test subjects, two targets need to be successfully tracked for the first 4 cases, and for the remaining 5 cases, there is only one target that needs to be tracked. The cases are as follows:

1. Target and normal person walk towards and then parallel to each other, with some partial overlap
2. Target and normal person walk across each other
3. Target and normal person walk towards each other, overlap and away from each other
4. Normal person surrounded by targets (1:2, 3 people in total)
5. Targets surrounded by (1:4, 5 people in total)
6. Child target in arms of a normal person
7. Target disappears from scene for 10s
8. Target disappears from scene for 30s
9. Target disappears from scene for 60s

The success of the system is measured as a boolean value (success or fail) based on its success in successfully matching the identity of the suspect to one single identity throughout the runtime of the system for that case. If the system matches the identity to the wrong person, even though it recovers later on, it will be counted as a failure.

3.3 Results and Analysis

The first metric that we were concerned with is the performance of the model on the training dataset. We check this to measure the performance of the model when it has domain-specific information, that is its' performance under ideal conditions. For our real system, the main result that we are concerned with is CMC-1 as for most cases, we don't have the labeled truth data so we have to assume that the top 1 model match is the correct match. Thus the CMC-1 accuracy signifies the expected model accuracy rate when it is used. However, the other metrics are included as a benchmark to give us more information on how the models performed. These 4 models were selected as some of the more well-known CNNs that are designed to be lighter or are just not a deep network due to our time limit constraint.

As can be seen from Table 3.1, the model based on ft-net-50 (ResNet-50) performs the best due to the general model complexity. However as the time required for the model to evaluate a certain ReID case is a different metric for us to optimize the system on, that affected our final choice for the model that we used for the system. The following is the average time the model spends processing a matching case when there is 1 target and about 3-7 images in the gallery set.

- ft-net-50 (ResNet-50): 19.69ms
- ft-net-18 (ResNet-18): 10.51ms
- mobilenet (Mobilenet): 18.19ms
- alexnet (Alexnet): 5.78 ms

From the list, we can see how AlexNet is about 3.4x faster compared to ResNet-50, however, this comes at the cost of losing around 13% accuracy. We eliminated mobilenet as it was only 1.08x faster with a 2% accuracy drop as we judged the improvement to be too marginal and would instead opt for ResNet-50. But what delivered the best mix of speed and accuracy is ResNet-18, delivering 1.87x faster speeds with a 6% accuracy drop on the training set. We further confirmed this result by sampling our recorded test cases data to create a ReID dataset which we used as an evaluation set to measure our model's performance on our real test cases. To do this, we selected 6 cases, 3 easier cases where the subjects are mostly visible and not overlapping each other, and 3 harder cases where there were more occlusions which affected the quality of the images in the gallery set. The frame by frame data was manually labeled and annotated to create the evaluation set that we used to measure the model's performance on our real-life data. The result of this evaluation can be seen in Figure 3.2, where we can see how the accuracy of the ResNet-18 model doesn't lose out to ResNet-50 (at worst performing 4% worse), and even performing better for half of the sampled cases, especially on Case 4.

Model Name	mAP	CMC-1	CMC-5	CMC-1
ft-net-50	0.776644	0.910036	0.969715	0.982185
ft-net-18	0.658812	0.853622	0.945962	0.96823
alexnet	0.542549	0.787411	0.913302	0.942696
mobilenet	0.703676	0.889846	0.955166	0.974169

Table 3.1: Performance of models on the test set of Market-1501 dataset after the training process, mAP (mean average precision) is the measured likelihood of all images of a person of the same identity being the top matches for the query, CMC-n measures the accuracy of a gallery match with the correct identity being in the top n matches

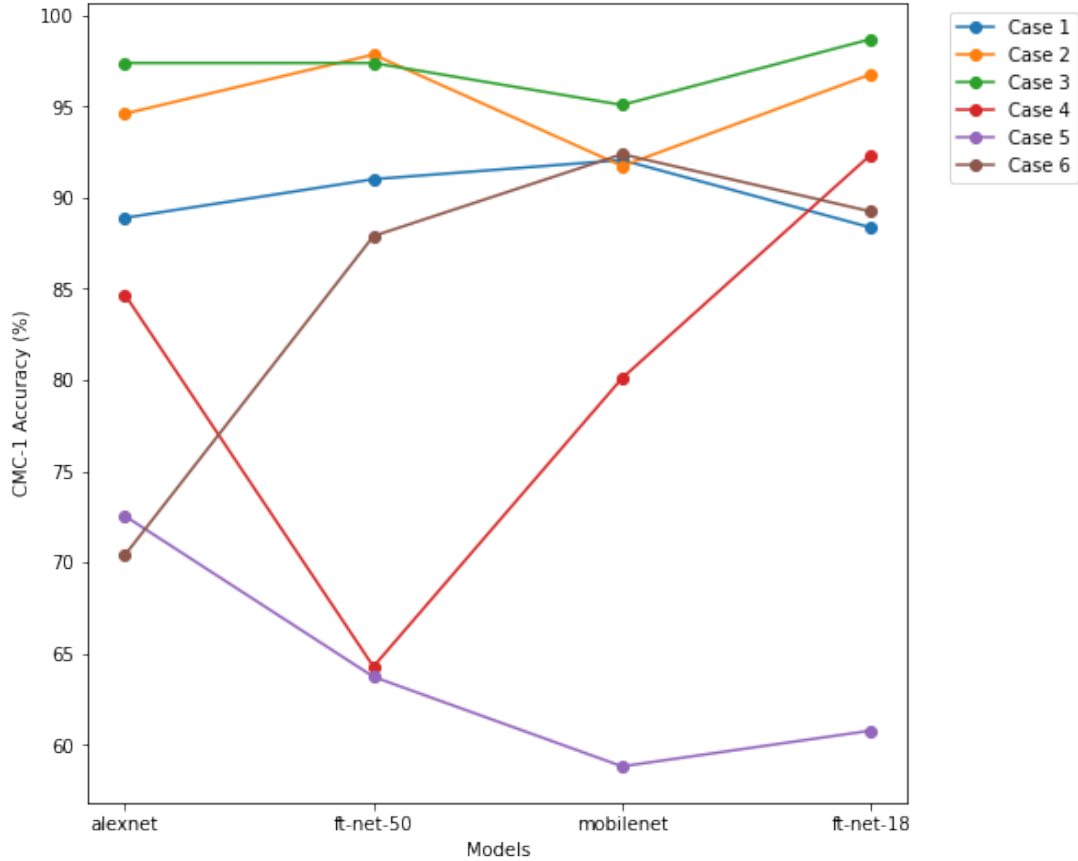


Figure 3.2: Result of testing our base models on our manually labeled personal dataset. CMC-1 accuracy, which measures the rate at which the model’s best gallery match’s identity is the same as the query. Case 1-3 are easy cases where subjects are mostly unobstructed and do not cross each other’s paths. Case 4-6 are hard cases where subjects often cross paths with one another, causing occlusions to happen.



Figure 3.3: Sample images from our manually labeled personal dataset. Top row: case 1-3, Bottom row: case 4-6.

As a result, we ran our simulated test cases using the ResNet-18 model as the backbone of the CNN module. However, even with the sampled cases and all our enhancements, the success rate of the whole module only amounted to 60% of all the test cases as can be seen in Table 3.2.

Case Description	Test Condition (D: distance between person in meters)			
	Normal Speed D > 1	Normal Speed D < 0.5	Fast Speed D > 1	Fast Speed D < 0.5
Target and normal person walk towards and then parallel to each other, with some partial overlap (\Downarrow)	✓	×	×	×
Target and normal person walk across each other (\times)	×	×	✓	×
Target and normal person walk towards each other, overlap and away from each other (\Uparrow)	✓	×	✓	✓
Normal person surrounded by targets (1:2)	×	×	N/A	N/A
Targets surrounded by normal people (1:4)	✓	✓	N/A	N/A
Child target in arms of a normal person	✓	✓	N/A	N/A
Target disappears from scene for 10 s	✓	✓	✓	✓
Target disappears from scene for 30 s	×	✓	✓	✓
Target disappears from scene for 60 s	×	✓	×	✓
Success Rate				60.00%

Table 3.2: Performance of the system on simulated crowd behavior cases under different test conditions. Speed denotes the walking speed of the subjects in the test case, N/A cases were not simulated.

3.4 Discussions and Conclusions

We implemented a real-time ReID system that can handle different crowd behavior under real-time and online constraints, achieving 60% accuracy on our test cases. The challenges with the system implementation and the steps taken to address them are detailed throughout the chapter. As is shown in Section 3.4, the ReID system performed rather poorly as compared to its' training data performance. We concluded that domain variance is a contributing factor. Similar to other deep learning applications, we assumed that if we were able to obtain annotated training data for our specific case to train our models with, then the performance would be comparable to the model's performance on the training data. But aside from this, we wondered as to why there is such a big difference in performance between the training set and our application. Because of this, we tried to sample data from the 3 easier cases and 3 harder cases to check the performance of the ReID model itself more quantitatively. As can be seen from Figure 3-2, the accuracy of the models dropped for cases 4-6 when compared to cases 1-3. This made us formulate the hypothesis that there is something about cases 4-6 that makes them very different from the training data that is provided to the models. We suspect that there must be a major contributing factor that is causing the difference in performance which we explore further in Chapter 4. We also concluded that the data in cases 1-3 are similar to the training data of the model, thus even with the domain variance, it is still able to accurately get the correct CMC-1 matches.

Chapter 4

Understanding the Cause of Failure in State-of-the-Art Person ReID Networks

Summary

In this chapter, we explored why our models in Chapter 3 had performed poorly and then extended the results to state-of-the-art models that are more complex to see if they were also plagued by the same issue.

4.1 Introduction

As was described in Chapter 3, there is a significant difference between model performance on the training dataset compared to its performance on the real system that we implemented. This difference hinted to us that there might be some assumption that the training dataset is making that is not aligned with the real situation that we are facing. Since with deep learning, models can only generalize patterns from things they have seen before, having a more accurate representation of the real situation in the training data often results in a performance increase. Furthermore, with all the different advances in the ReID field, we wanted to see how the best models in the field perform and whether these approaches would be able to tackle the challenges that we are facing.

Thus, we propose a benchmark that we utilized to select a few key models that represented the best performing model in their category. We then used these models to perform three experiments that were designed on top of each other. As we understood more about what the models are doing, we designed the next experiment to further explore the factors that are significant to the model performance.

4.2 Methods

In this section, we detail the implementation and rationale behind our benchmark, and the 3 experiments that we used to measure how the models are performing.

4.2.1 Initial Benchmark

To begin our experiments, we wanted to compare the performance of the models that we have against the performance of state-of-the-art (SOTA) approaches to see if a CNN architecture change would allow us to overcome the challenges that we face in Chapter 3. There are 2 metrics that we considered as key constraints for developing the ReID system, they are accuracy and time. To replicate this, we used the mAP metric as a measure for how well a model performs. The benchmark is run on Market1501's evaluation set as a comparative benchmark. The models are also all trained using Market1501.

First we selected models based on the 3 categories of ReID approaches that we established in Section 2.1. We picked out Bag of Tricks [10], OSNet [12], PLR-OSNet [13], DGNet [16], LightMBN [15], TopDBNet [17], and ABDNet [14] as they were all part of the highest performing models on the Market1501 benchmark. We included the models that we used in Chapter 3 as a baseline comparison to see how these SOTA models stack up against our existing models.

There are 2 benchmark actions that we performed, the first is to check whether batch sizes had any impact on the performance of the model. Batch size is the number of images that are stacked together before being sent to the model to reduce GPU copying overhead. This is done to see whether there is an optimum or limit to the batch size before the performance of the model drops. Since batch sizes are usually checked in multiples of 2, we tested batch sizes from 8 to 512 and multiplied the batch size by 2 with each step. The second benchmark is to measure the GPU inference time on fixed batch size, we picked 128 as the batch size as that number is the expected number of images that our system in Chapter 3 needs to infer every round. For both benchmarks, we simply let the model run its inference and then measure the related variables as they are returned by the model. We did not include the CPU comparison time that is required to rank the similarity of the features like the gallery size for the test set is much larger than our expected use case.

4.2.2 Experiment 1: Examining Model Failure Cases

After running the benchmark, 4 main models were selected as the best performing models based on the results that are shown in Figure 4.2. These models are Bag Of Tricks, PLR-OSNet, LightMBN, TopDBNet and the reason why they are selected is explained in Section 4.3. With these 4 models, we performed the following steps:

1. Run inference on Market1501 evaluation set using one of the models.
2. Record the failure cases where query and gallery matched to different identities based on CMC-n metric.
3. Repeat steps 1 and 2 for all models.

4. Construct intersection sets between the failure case sets for each model.

The intersection set is used for us to identify the most difficult cases that all SOTA models are unable to correctly match so we can study the reason behind these failures. We performed 2 variations of the steps, where we changed whether the same camera matches are allowed or not. In regular ReID, the same camera matches are restricted thus only allowing matches between different cameras. We wanted to observe the effect of enabling or disabling this setting to check whether the cases are easier when the viewpoints of the camera are the same.

For each model, we also constructed a feature visualizer to visualize the image regions that create activations in the generated feature map. This is done by taking the outputted feature maps from the model at the final convolution layer, stacking them all together, and rescaling it to fit the original image dimensions (256x128). We performed this on all the images in the CMC-1 failure intersection set to see what the model was paying attention to.

4.2.3 Experiment 2: Exploring the Impact of Occlusion

This experiment is designed to verify the impact of occlusion on the performance of the models. To do this, we performed 2 tests on 2 evaluation datasets that contain occluded cases. The first test is performed on our manually annotated site dataset from Chapter 3. These are the steps of the experiment:

1. Run inference on the personal dataset using one of the models.
2. Record the CMC-1 accuracy of the model.
3. Repeat from step 1 with the next model.

The second test is conducted similarly to the first one, it is carried out on the OccludedReid dataset [28]. The following are the steps of the experiment:

1. Select a query and gallery setting
2. Select a model and run inference on the OccludedReid dataset based on the query and gallery settings.
3. Record the CMC-1 accuracy of the model.
4. Repeat step 2-3 for all the models.
5. After step 4 is done, start from step 1 for the next gallery setting.

To construct the query and gallery sets, we select the 1 whole image and 1 occluded image for each identity as the query set for the whole and occluded setting respectively. The remaining 4 whole and 4 occluded images are both used as the gallery set. Resulting in 200 images in the query set and 800 images in the gallery set for the whole and occluded settings. Changing the query and gallery settings means changing the combination of query and gallery sets.

4.2.4 Experiment 3: Exploring Whole and Occluded Images and Their Impact on Model Performance

For experiment 3, we explored the effects of adjusting the ratio of whole and occluded images on the model performance. We ran two tests in this experiment, both were conducted on the OccludedReid dataset and we used the same whole and occluded splits like the one in Experiment 2. For the first test, we mixed the whole and occluded images in the gallery set. For example, a 1:3 occluded:whole image ratio means that 1/4th of the occluded gallery and 3/4th of the whole gallery make up the final gallery set on which we run the model inference. The steps of the test are as follows:

1. Select a query setting, either whole or occluded.
2. Select a gallery ratio and construct the gallery set.
3. Select a model and run inference on the OccludedReid dataset based on the query and gallery settings.
4. Record the CMC-1 accuracy of the model.
5. Repeat steps 3-4 for all the models.
6. Once step 5 is finished, start from step 2 with a different gallery ratio.
7. Once step 6 is finished, start from step 1 with the other query setting.

For the second test, we wanted to see the effect of adding images of a different type to the whole gallery dataset to see if there was a bias to certain parts of the dataset. To do this, we progressively add images from the other gallery setting, for example, if the base gallery type is whole, then we will be adding the occluded images to the gallery set. The following are the steps of the experiment:

1. Select a query setting from whole or occluded.
2. Select a base gallery setting from whole or occluded.
3. Add a single image for each identity from the opposite gallery setting to the gallery set.

4. Select a model and run the inference on the query and gallery setting.
5. Record the CMC-1 accuracy of the model.
6. Repeat from step 4 for all models.
7. After step 6 is finished, repeat from step 3 and add another image to the gallery set.
8. After step 7 is finished, repeat from step 2 and change the base gallery setting.
9. After step 8 is finished, repeat from step 1 and change the query setting.

4.3 Results and Analysis

4.3.1 Benchmark Results

From the benchmark experiments results in Figure 4.1, we observe that the batch size has no real effect on the mAP except for the ABDNet model. Since for all the other models we didn't see any issues, we pinpointed the issue to be in the architecture of how the model itself does feature extraction. For most models, this is not an issue, however, for specific models, there might be some batch sizes that directly impact its' performance. Thus this test is key in selecting a standard batch size if there is a need to perform inference on different batch sizes as it may affect the model performance.

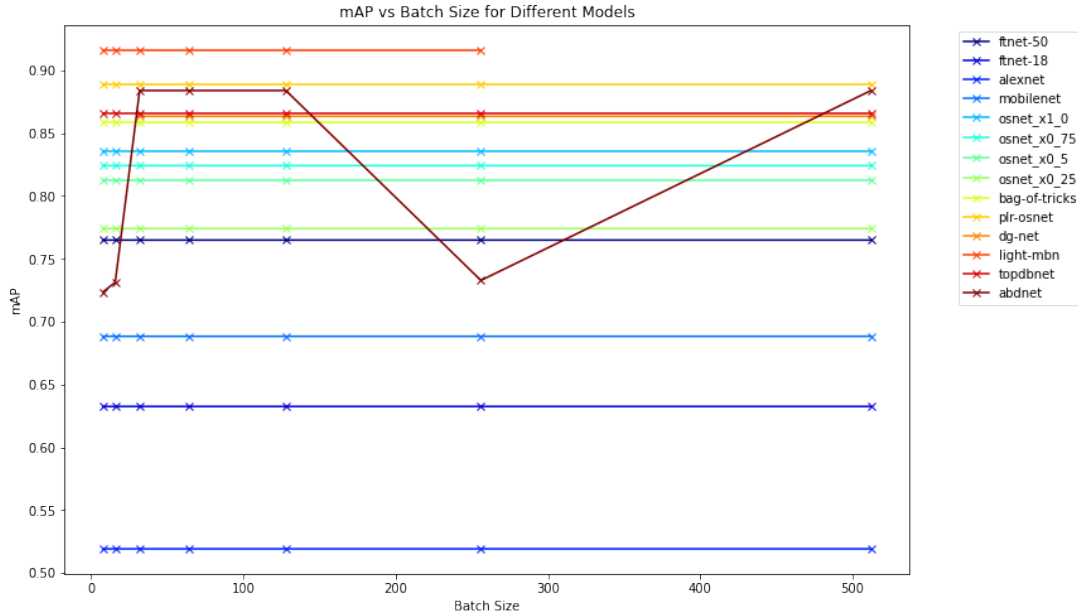


Figure 4.1: Effect of varying inference batch size on model mAP. mAP is a metric of ReID model performance that measures its precision in fetching images of the same identity

In Figure 4.2, there is a general trendline of the models keeping the model inference time to be under 200ms overall. However, with most SOTA approaches that are newer and

have better performance, we see that time is not something that they optimize on as we see a general trend of the GPU inference time taking longer. Seeing that GPU inference time is directly correlated to the complexity of the model and the fact that there is not much research that is done to make the inference time faster, this becomes a bottleneck if we were to design a real-time system. For a lot of the newer models, they seem to trade off a large amount of inference time for a minimal amount of accuracy increase. For example, between PLR-OSnet and LightMBN we see an increase of 3x in terms of time taken for around 4% extra mAP. Comparing this to the performance increase between AlexNet and PLR-OSNet where PLR-OSNet also needs a similar 3x longer time, but it is traded off for around 34% mAP. We identified that the optimum models exist in the bottom right part of Figure 4.2 where there is a mix of high accuracy but with minimal cost in inference times.

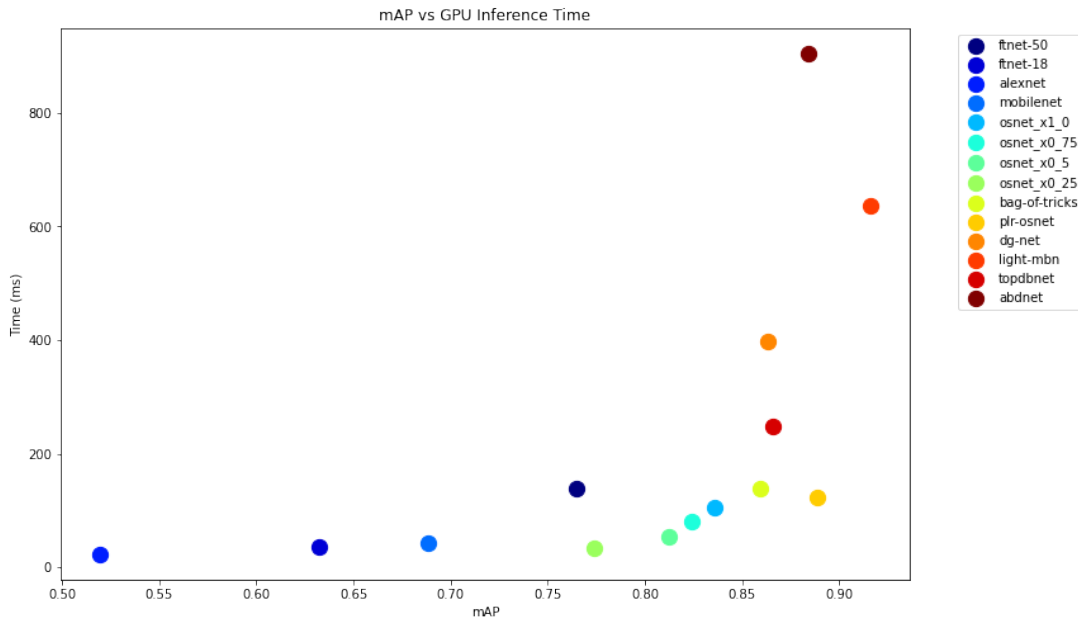


Figure 4.2: mAP and speed of inference for each model with a batch size of 128. mAP is a metric of model performance

Due to this, we referred back to the 3 categories to the approaches we defined in Section 2.1 and the top-performing model from each category. This resulted in our choices for Bag of Tricks for the global feature focus, PLR-OSNet for the local feature focus, and TopDBNet for the auxiliary features focus. Out of these 3, we identified that the local feature focus models make up most of the models that we test on, as such, we included the 2nd best model which is LightMBN to see how a more complex model performs in our test cases. These 4 models are the models that we use for further tests on the 3 experiments.

4.3.2 Experiment 1 Results

We opted to observe the CMC-n metric to judge the top match results of the models on their training dataset. This was done to understand what are the marginal and difficult cases that all of the newest SOTA approaches are unable to solve. Figures 4.3-4.5 contain the CMC-n intersection sets that show to us the total number of failure cases and the sizes of each intersection sets. Out of 3368 queries, we can see how the SOTA models perform well that even under the strictest metric (CMC-1) there are only 55 failure cases where all 4 models failed to meet the criteria. These 55 failure cases are the ones that interest us, and we wanted to see whether there was a common cause of failure that caused these.

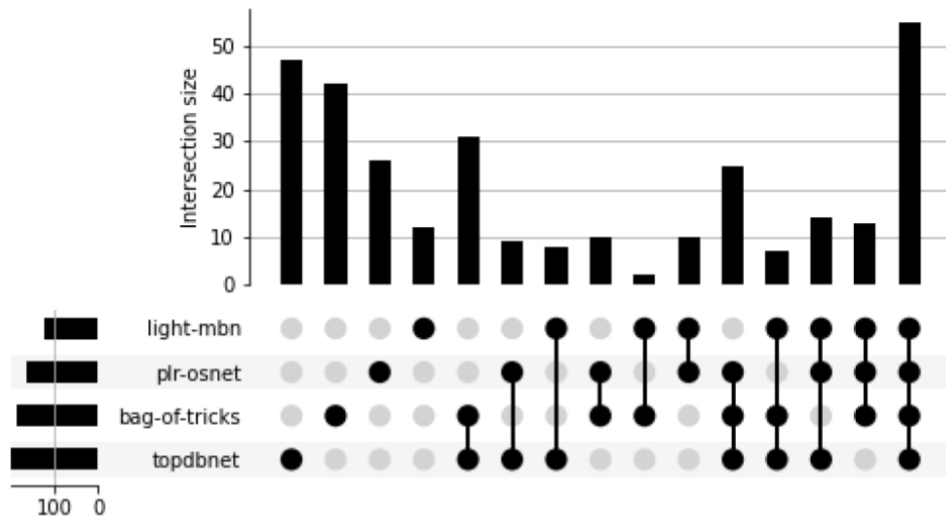


Figure 4.3: Size of intersection sets of ReID failure cases between the models with the CMC-1 metric. Intersection sets contain the images where all the models listed are unable to meet the CMC criteria. CMC-n checks for whether the correct identity match is within the top n matches. Size of each model’s failure set is shown on the left bar.

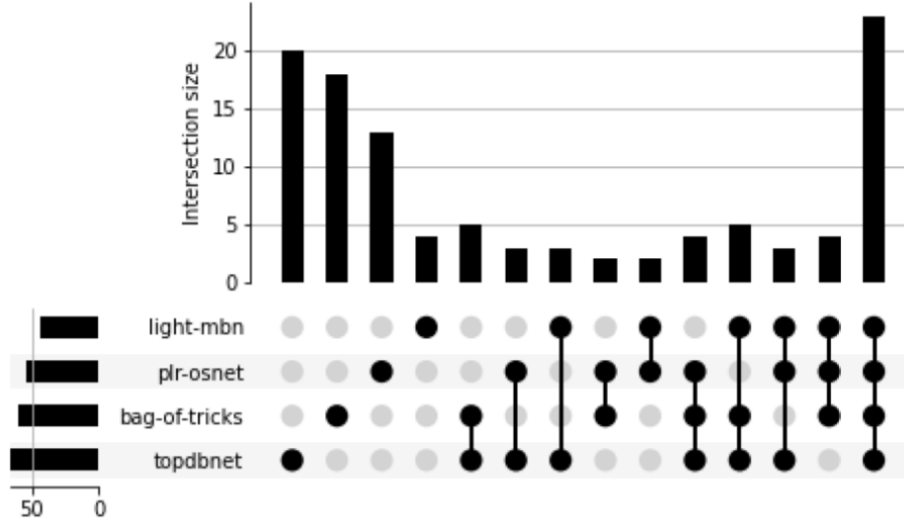


Figure 4.4: Size of intersection sets of ReID failure cases between the models with the CMC-5 metric. Intersection sets contain the images where all the models listed are unable to meet the CMC criteria. CMC-n checks for whether the correct identity match is within the top n matches. Size of each model’s failure set is shown on the left bar.

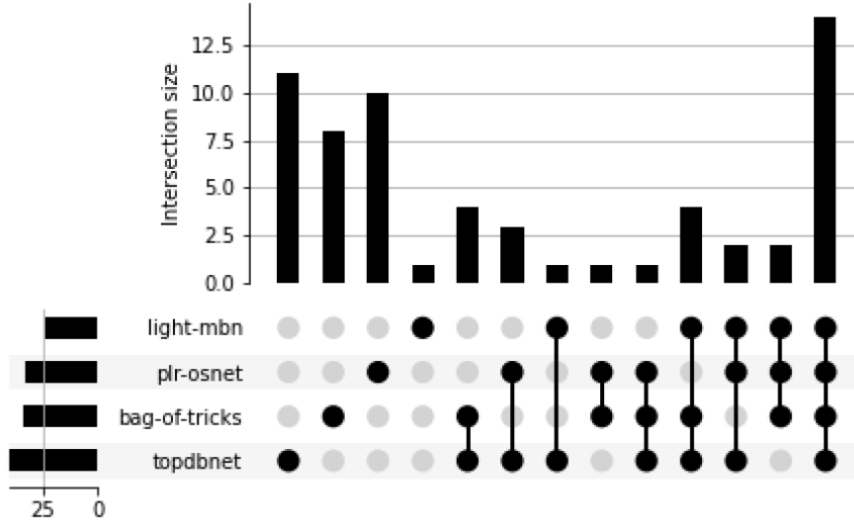


Figure 4.5: Size of intersection sets of ReID failure cases between the models with the CMC-10 metric. Intersection sets contain the images where all the models listed are unable to meet the CMC criteria. CMC-n checks for whether the correct identity match is within the top n matches. Size of each model’s failure set is shown on the left bar.

From the 2nd variant of the test where we allowed the same camera matches we discovered that this trivializes most of the problematic matches. This is caused by a similar viewpoint that eliminates most of the challenges in matching the identity of the person. As can be seen in Figures 4.6-4.8, the intersection set sizes are very small meaning that models have 99% accuracy. This guarantees the ability of these models if they were to be used in the same camera tracking as it seems that the models are very capable of matching identities in this case.

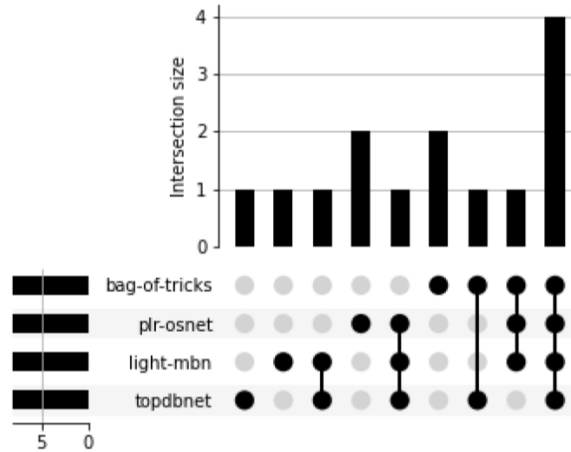


Figure 4.6: Size of intersection sets of ReID failure cases between the models with the CMC-1 metric when same camera matches are allowed. Size of each model’s failure set is shown on the left bar.

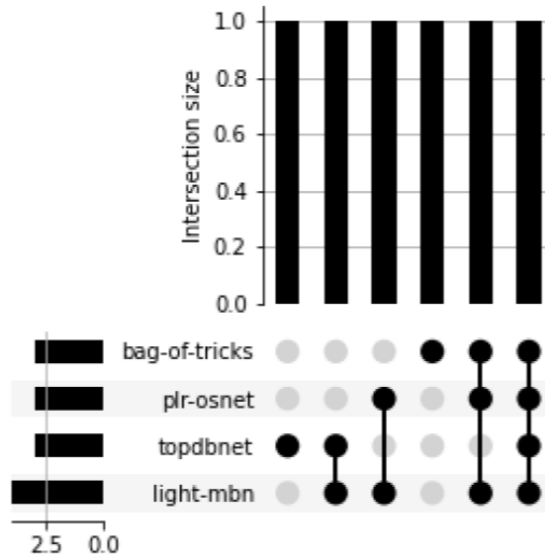


Figure 4.7: Size of intersection sets of ReID failure cases between the models with the CMC-5 metric when same camera matches are allowed. Size of each model’s failure set is shown on the left bar.

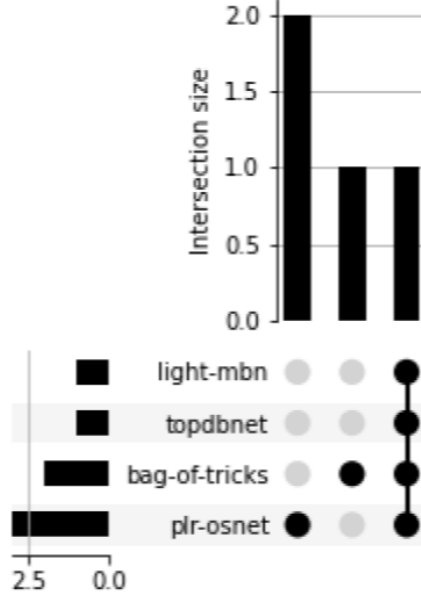


Figure 4.8: Size of intersection sets of ReID failure cases between the models with the CMC-10 metric when same camera matches are allowed. Size of each model’s failure set is shown on the left bar.

As the 2nd variant of the test trivializes the Person ReID problem, we focused on the results from the 1st variant where same camera matches are not allowed. We chose the results from the strictest metric which is CMC-1 as we concluded that the results from the larger CMC-n tests would only be a subset of the failure cases for CMC-1. So we gathered all the images from the 55 failure cases (Figure 4.9) and qualitatively checked the results of each of those failure cases. In doing so, we discovered that 30 out of the 55 are failures caused by mislabelling or duplicate data. As a human observer that would look at the images of the results, it can be concluded that the images are of the same person. Thus, the actual failure case set only consisted of 25 out of 3368 queries, which means that the SOTA models are only unable to handle 0.74% of all the query cases.

In these 25 cases, we qualitatively looked through the results of the match and ran the feature visualizer on them to see what the models were paying attention to. A sample result can be seen in Figure 4.10 and from this process, we concluded that for the failure cases, occlusion was a core factor that appeared in most of the images. As we looked at the regions where the models paid attention, we noticed that occlusion causes the models to be unable to do a 1 to 1 comparison on the resulting feature map. Also, as we did a breakdown of the training data, we realized that most of the images in the training dataset only consist of clean, whole images of each identity. We made the hypothesis that occlusion is the main cause of the failures of these models. Experiment 2 is then carried out to verify these results.



Figure 4.9: The query images in the CMC-1 failure cases intersection set of the 4 selected models.

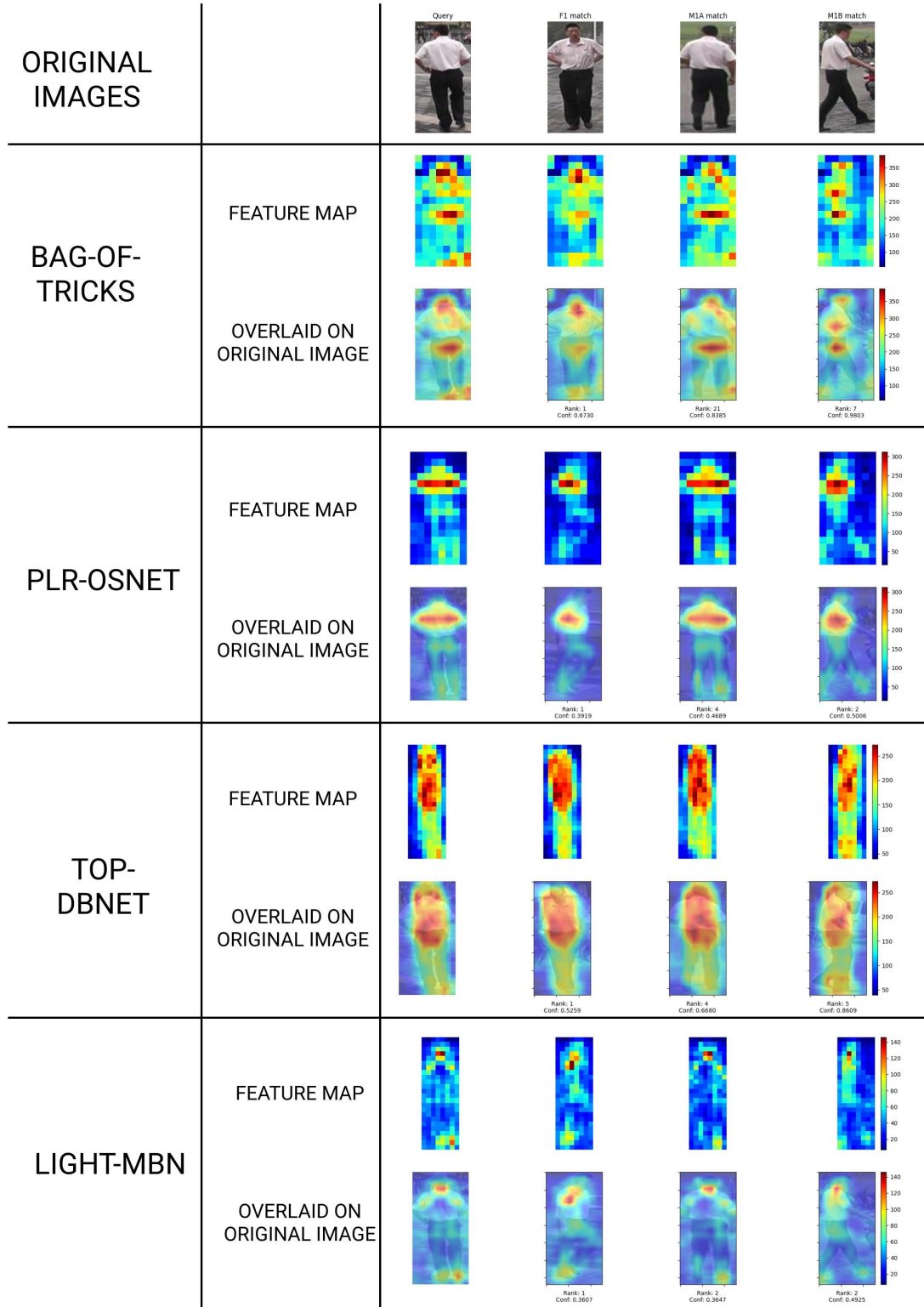


Figure 4.10: Sample results from the Feature Visualizer on the regions that the 4 selected models pay attention to when creating the image's feature representation

4.3.3 Experiment 2 Results

To verify whether occlusion is indeed a major cause of failure in the models, we used the same 4 models and added the ResNet-18 model as a baseline for comparison for the tests in Experiment 2. For the results from the first test on the personal dataset (Figure 4.11), looking at the overall accuracy, all 5 models perform rather well, the average CMC-1 accuracy overall cases range from 89.5%-94.7%. However, we observed that the average model accuracy has a big gap between the results for Case 1-3 as compared to the results for Case 4-6. The average CMC-1 accuracy for Case 1-3 would lie in the range of 94.7%-98.5% whereas, for Case 4-6, they are in the 75.0%-82.5%. We realized that in cases 4-6, there are more occluded samples in the gallery which agrees with our hypothesis that occluded images will greatly affect the performance of the model. The only outlier, in this case, was the TopDBNet model which practically achieved similar accuracy for all 6 cases and we attribute this to its network architecture design. Since it was designed to look at auxiliary information sources, it is not too affected by the partial loss of information that is caused by occlusion. Such an effect is not observed on all the other models that have a stronger focus on the information that it specifically learned from the training dataset. This can be seen in how the best model, PLR-OSNet has the worst average performance on cases 4-6.

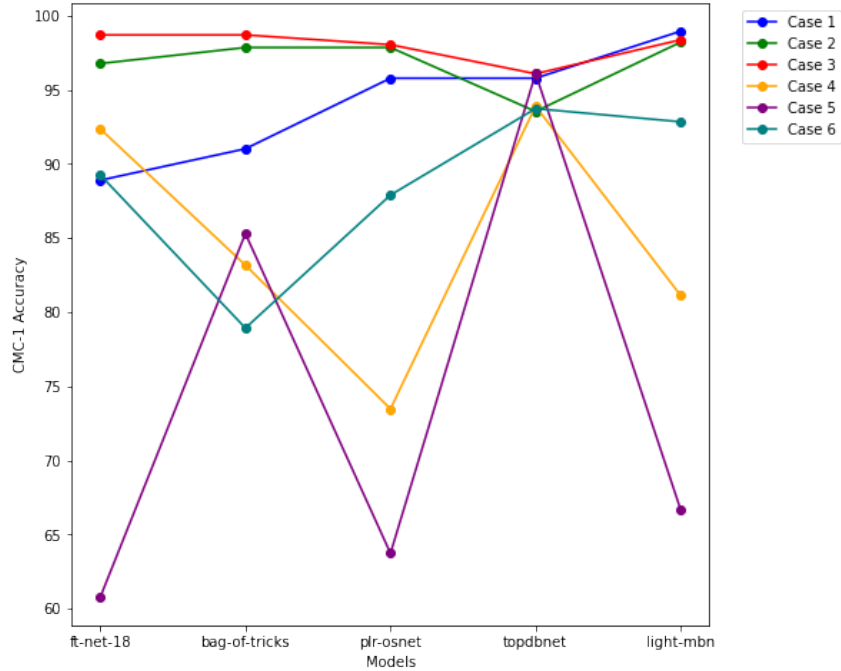


Figure 4.11: Result of testing SOTA models and ResNet-18 as a benchmark on our manually labeled personal dataset. CMC-1 accuracy measures the rate at which the model’s best gallery match’s identity is the same as the query. Case 1-3 are easy cases where subjects are mostly unobstructed and do not cross each other’s paths. Case 4-6 are hard cases where subjects often cross paths with one another, causing occlusions to happen.

To further verify our hypothesis, we ran the 2nd test on the OccludedReID dataset which is an evaluation set that contained both whole and occluded samples for each identity. As we varied the whole and occluded setting of the query and gallery, a clear downtrend is observed that shows that for all 5 models, the accuracy is the lowest when the models have to deal with an occluded query and occluded gallery set. The accuracy difference is massive as can be seen where it is all above 91% for the whole to whole case, while for the occluded to occluded cases the range is 25%-52.5%.

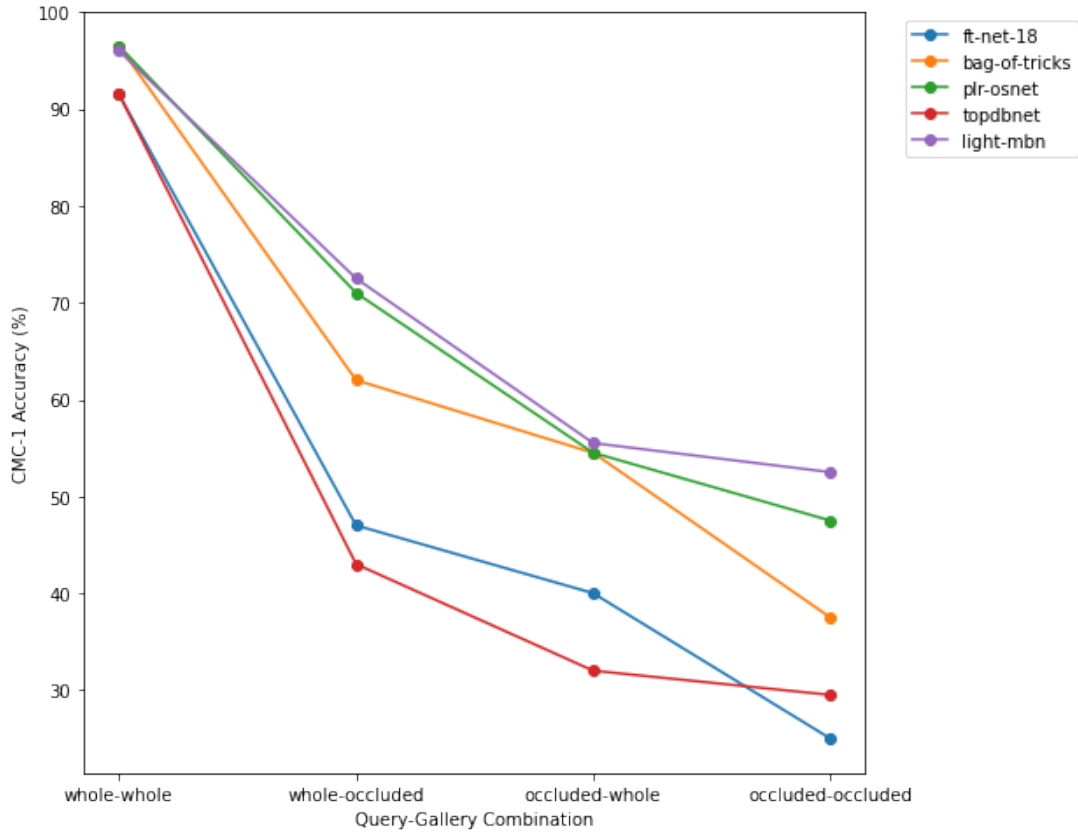


Figure 4.12: Result of testing SOTA models and ResNet-18 as a benchmark on the OccludedReid dataset. The contents of the query and gallery set are varied according to the settings in the Y axis to measure the effect of occlusion. CMC-1 accuracy measures the rate at which the model’s best gallery match’s identity is the same as the query.

4.3.4 Experiment 3 Results

For this experiment, we wanted to further explore the interplay between whole and occluded images. Once again we use CMC-1 accuracy as a metric of model performance. From the results of the first test (Figure 4.13), we observed a clear separation between the results when the query is whole as compared to when the query was occluded. Looking at the graph trends, we recognized that there is a downtrend for both query settings as the number of occluded images in the gallery set grew. However, for the whole query setting, there is a sharp drop in accuracy (on average a 29.9% accuracy drop) when the gallery

only consists of occluded images. This finding suggests that the models are overfitting to the whole cases and are unable to generalize well enough to handle the occluded cases.

Another observation that we made was the increase of accuracy when comparing the whole query setting with the occluded query setting, even when the gallery set was fully occluded. We can see on average a 20.7% accuracy difference between the two settings which highlights the importance of having whole image queries as this greatly affected the model performance when handling occluded images. Overall, the average accuracy of the whole query setting is 83.02% whereas, for the occluded query setting, it is merely 43.54%, a sharp 39.48% drop.

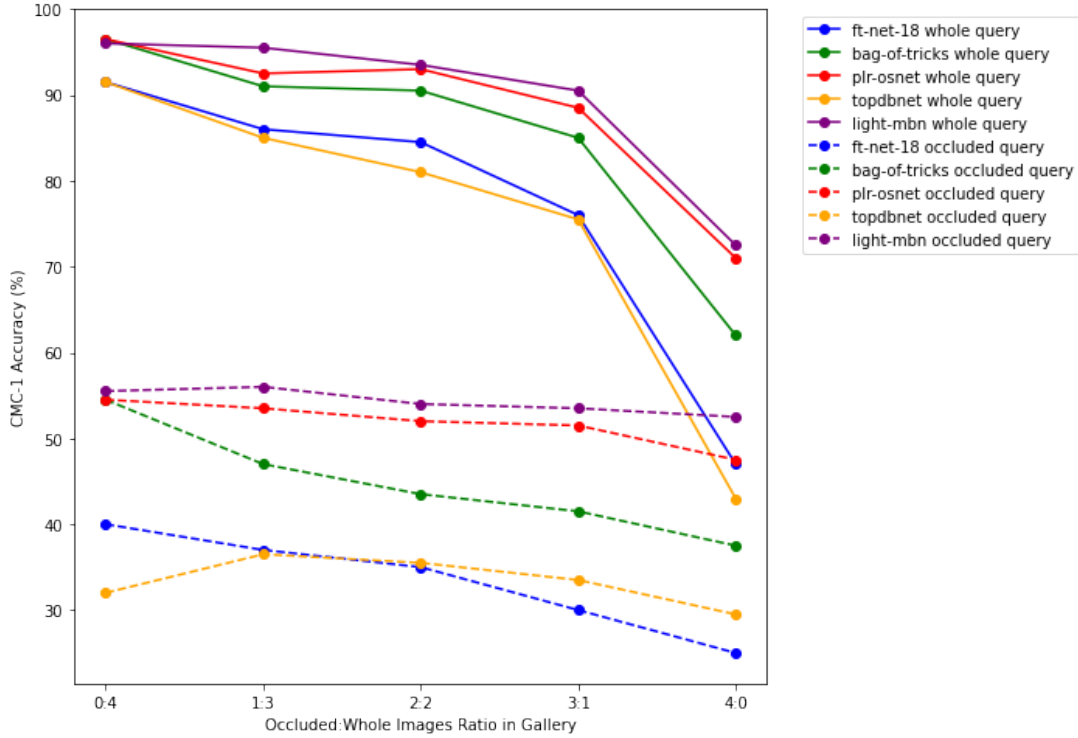


Figure 4.13: Effect of changing the ratio of occluded and whole images in the gallery set on model’s performance on OccludedReID Dataset with different query types

From the results of our second test (Table 4.1 and 4.2) we see no new surprises as compared to the results from the first test. It serves to verify the trends that were present in the first test regarding how a model’s accuracy is greatly affected by the quality of the query image and the number of whole images of the person in the gallery set.

Query Type	Base Gallery Type	Extra Gallery Images	ft-net-18	bag-of-tricks	plr-osnet	topdbnet	light-mbn
Whole	Whole	None	91.5	96.5	96.5	91.5	96
		1/4 Occluded	91.5	96	96	91	96
		2/4 Occluded	92	96.5	97	90.5	96.5
		3/4 Occluded	92	96	96.5	90.5	97
		All Occluded	92	96	96.5	90.5	97
Whole	Occluded	None	47	62	71	43	72.5
		1/4 Whole	85	92	95	84.5	93.5
		2/4 Whole	89.5	94	96.5	88.5	94
		3/4 Whole	91	95.5	96.5	89	96
		All Whole	92	96.5	96	90.5	97

Table 4.1: Effect of adding extra images to the gallery set on model performance with whole query images, accuracy drops when there are no whole image samples in the gallery. The numbers in the table are the percentage CMC-1 accuracies of the models. CMC-1 accuracy denotes the model’s rate of having the best gallery match as the true identity match with the query.

Query Type	Base Gallery Type	Extra Gallery Images	ft-net-18	bag-of-tricks	plr-osnet	topdbnet	light-mbn
Occluded	Whole	None	40	54.5	54.5	32	55.5
		1/4 Occluded	40	48.5	55	37	57
		2/4 Occluded	38	46.5	54	36.5	56
		3/4 Occluded	37	47	55	36.5	58
		All Occluded	38.5	48	55	37	60
Occluded	Occluded	None	25	37.5	47.5	29.5	52.5
		1/4 Whole	34.5	43	52.5	34	57.5
		2/4 Whole	36.5	43.5	52.5	36.5	58
		3/4 Whole	37	45	52.5	36	58
		All Whole	38.5	48	55	37	60

Table 4.2: Effect of adding extra images to the gallery set on model performance with occluded query images. The numbers in the table are the percentage CMC-1 accuracies of the models. CMC-1 accuracy denotes the model’s rate of having the best gallery match as the true identity match with the query. Note the low accuracy when compared to the results in Table 4.1

4.4 Discussions and Conclusions

In this chapter, we established a benchmark and a set of experiments that can be used to evaluate model performances between whole image and occluded image cases. From those experiments, we concluded that occlusion is a major factor behind the failure cases of the models. When dealing with occluded cases, model performance would see extreme drops in the range of 38.5%-66% when compared to the general case. Furthermore, most SOTA approaches are not built to handle these occluded models and they perform rather poorly in general when faced with heavier occlusion.

We also discovered that having good quality query images is much more important than ensuring that all the images are whole. With whole query images, model accuracy shows great improvements over when the model only has occluded query images to work with. This makes this an important design consideration when developing systems in the future that need to perform ReID when occlusions happen.

Finally, we hypothesize that there is a bias in the training data that improperly models

the real-world situation. We believe that the Market1501 training data is heavily biased to whole images, causing models trained on this dataset to have poor performance when dealing with occluded cases. To check whether this hypothesis is true, we attempted more experiments that are discussed in Chapter 5.

Chapter 5

Impact of Occluded Training Samples in Occlusion Situations

Summary

In this chapter, we verify the effect of substituting the training data with data that contain more occluded samples. We also explore a better representation of the evaluation set which can better model the expected model performance in real situations.

5.1 Introduction

In Chapter 4, we observed how occlusion greatly impacts the model’s accuracy when matching identities. A hypothesis that was formed is that this was caused by the training dataset that was biased towards whole images, thus causing the models to be unable to learn patterns that are present in real-life cases. To test this assumption, in this chapter we explore the use of a dataset with occluded samples in its training data to see whether occlusion has a pattern that can be learned using CNNs.

5.2 Methods

In this section, we detail how we perform the training data replacement and how the test to measure the impact of the model was performed.

5.2.1 Dataset and Models

The training data that is used for this experiment is the Occluded-DukeMTMC Dataset which we picked because it has 9%/100%/10% occluded images in the training, query, and gallery set respectively [26]. This dataset consists of a large number of samples that would make it comparable to Market1501 in terms of the amount of training data. Other datasets like OccludedReid were intended to be evaluation sets instead of training sets hence why we did not select them for this experiment.

The models that were trained using the Occluded-DukeMTMC dataset are the same 4 SOTA models that we chose in chapter 4. We also retrained the ResNet-18 model using the dataset as a benchmark for comparison. The training process follows the standard process for training ReID models that are used by [9]. After the training process was

finished, we evaluated the models by using them to run inference on the query and gallery set and recorded the CMC-1 and mAP as metrics for model performance.

5.2.2 Experiment Setup

Two tests were done to measure the capability of the model in solving occluded cases. The experiment is run using the same settings as the two tests defined in Section 4.2.3 and the results of this experiment are compared to the results of the experiment in Section 4.2.3 to measure the difference in performance.

5.3 Results and Analysis

The query and gallery set that is used for evaluation on the Occluded-DukeMTMC dataset consists fully of occluded cases. This results in the evaluation set itself becoming a test of the model’s performance on occluded cases. As is shown in Table 5.1, LightMBN which is the most complex model has the best performance on the evaluation set. This suggests that the part-based models can discover patterns that help them with dealing with the occlusion cases. This suggests the existence of patterns that are discoverable by the models to help them deal with occluded cases. Another thing that can be seen is that TopDBNet didn’t perform that well compared to the models that are more focused on the parts. This is likely due to the network design that focuses on learning more general patterns, when there are occluded data in the training set it might cause such data to be interpreted as noise.

Models	CMC-1	mAP
Resnet18	26.06	16.29
bag-of-tricks	49.5	42.8
PLR-OSNet	61	53.2
TopDBNet	52.4	42.7
LightMBN	64.98	58.44

Table 5.1: Performance of selected models on the test set of OccludedDuke dataset after the training process, the mAP and CMC-1 are metrics for model performance.

To further verify the results of the training, we compare the results of the models that are trained using Occluded-DukeMTMC with the models that are trained using Market1501. Averaging the difference of CMC-1 accuracy of the models, we see an increase of 4.48% in the CMC-1 accuracy. However, if we exclude ResNet-18, for all the SOTA models, the increase goes up to 6.72%. Furthermore, when we examine the average increase for the whole-occluded / occluded-whole / occluded-occluded settings, we see an on average accuracy increase of 3.8%/4%/10.9% respectively. This reinforces the hypothesis

that the models that are trained on the Occluded-DukeMTMC which contain occluded samples are better at handling occlusion cases.

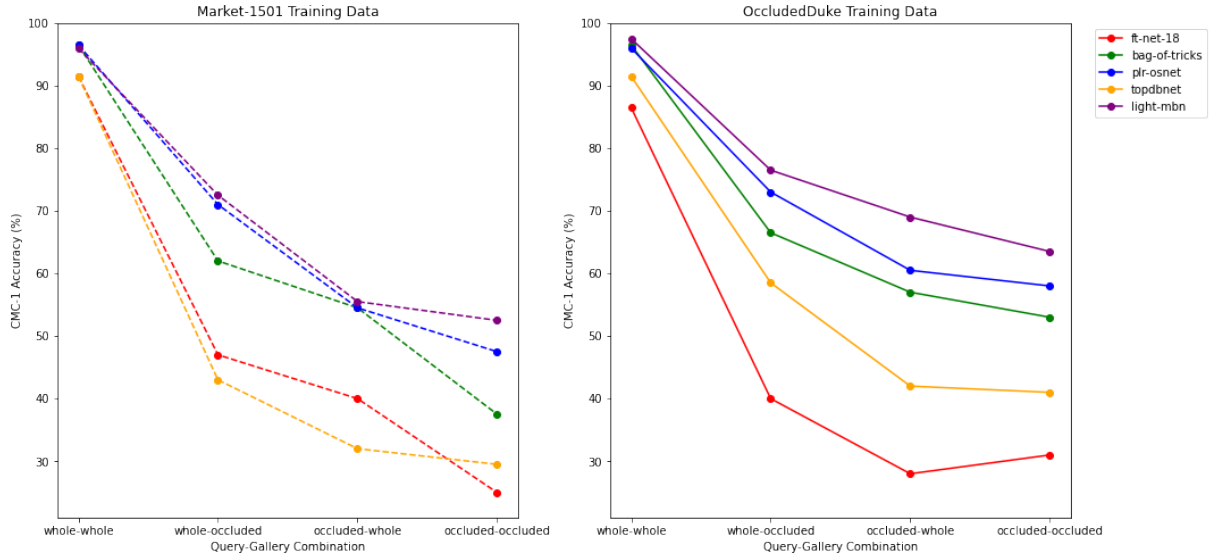


Figure 5.1: Comparing the effects of having occluded training data to the model’s CMC-1 accuracy on the OccludedReid test set. CMC-1 is a metric for a ReID model’s performance.

Our last test verifies the capability of the newly trained models on our real-life use case. We benchmarked it on our dataset and we observed a sizable increase in accuracy for the SOTA models, except for TopDBNet. The average accuracy increase for cases 4-6 for Bag Of Tricks / PLR-OSNet / LightMBN is 8.37%/17.57%/13.18%. Even ResNet-18 reported an average 2.81% accuracy increase. TopDBNet experienced a decrease of -9.76% to its average accuracy, and the cause is likely directly related to its performance on the training dataset. Furthermore, some models namely PLR-OSNet and LightMBN experienced a decrease in average accuracy for cases 1-3 of -0.66% and -2.35%. This is possible because some of the features that become more important when dealing with occluded cases might have caused the model to look at things more generally. Thus small important details might have been deemed as less important features, causing the whole case to suffer a little. However, for all models except TopDBNet, there is an overall increase in average model accuracy.

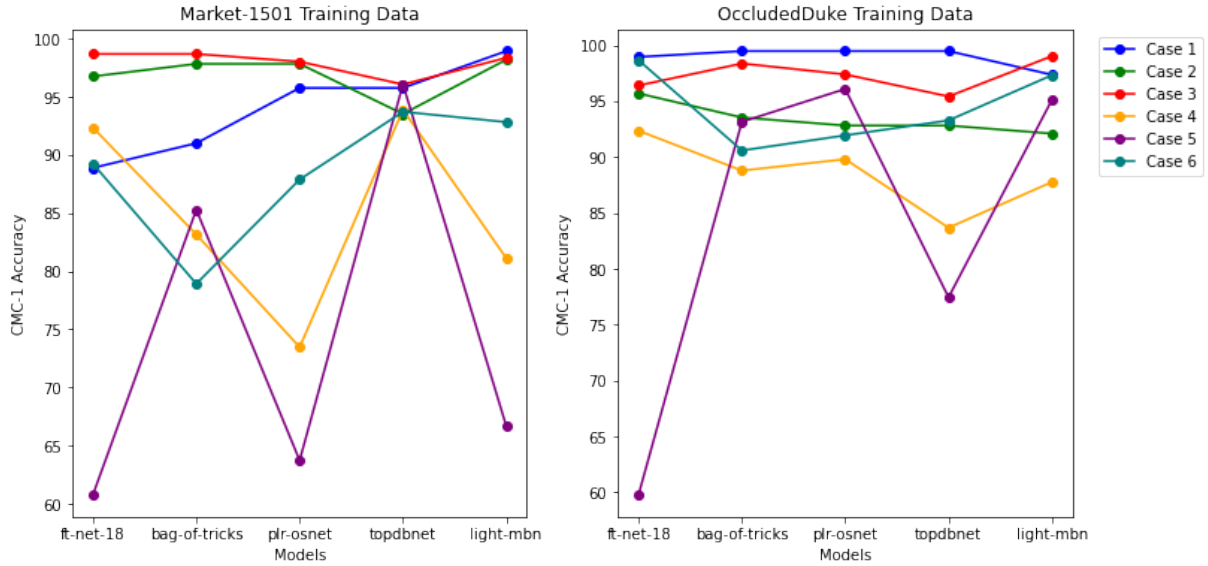


Figure 5.2: Comparing the effects of having occluded training data to the model’s CMC-1 accuracy on the Personal dataset. CMC-1 is a metric for a ReID model’s performance.

5.4 Discussions and Conclusions

We conclude that substituting the training data with the Occluded-DukeMTMC positively impacted the model’s ability to handle occluded cases. This was evident in the evaluation results on the three evaluation datasets that we have. However, we observed that the improvement was not consistent in all the models. Primarily, we saw no improvement and even worse performance with the newly trained dataset on the TopDBNet and our baseline Resnet-18 model. For TopDBNet, we suspect that this is caused by the design of the network that focuses on learning more generalized features. This allowed it to pay attention to other features that help it classify the different identities better under occlusion. However, when they encountered the same data in the training data, the occluded data likely became noise to the learning process. Making it harder for the model to find an optimal set of features. Thus causing the model’s lower overall performance. compared to when its initial training data was cleaner. For ResNet-18 we suspect that the model was simply not complex enough to capture the patterns that were present in the new training data we provided it. Resulting in an overall decrease in performance as the new sample merely caused it to be more confused.

Though this method doesn’t work for all model architectures, especially those that already generalize the learned features by design. Furthermore, there is a small trade-off in the accuracy of the models for the whole case. Future research on this topic can explore the possibility of training using a fully occluded dataset and the effect that this has on the ability of the model’s capabilities in handling both occluded and whole cases. Additionally, a study on the impact of different network designs or structures and training processes on the performance of the model could also provide some insights into the effects

of the design on its performance on whole and occluded images.

Chapter 6

Conclusions, Limitations and Future Works

6.1 Conclusions

In this thesis, we proposed and implemented an online open-world person ReID system in a screening system that monitors a crowd using algorithms and deep learning models. We explored the challenges with creating such a system and used preprocessing and post-processing techniques to address those challenges and optimize the system performance. This system passed 60% of our test cases. Even with all this, we discovered a discrepancy between the expected performance of the model from its performance on the training set as compared to its performance in the real situation.

As such we investigated the main causes of failure for not only the model that we used for our system but also for SOTA approaches. Through this, it was uncovered that occlusion was a major cause of failure, causing on average 16.0%-19.7% accuracy drop on our personal dataset and 38.5%-66% accuracy drop when dealing with occlusion cases on OccludedReid dataset. We observed that for the SOTA models, accurate predictions can be greatly affected by the quality of the query image, occluded query images cause the identity match to have much worse performance compared to whole query images.

We also discovered that using training data that contain occluded samples has a positive effect on models that are more complex and able to look for deeper patterns in the training data, causing accuracy increases in the range of 8.37%-17.57% on our personal dataset and on average a 6.72% accuracy increase for the SOTA models on OccludedReid dataset. However, we observed that it has a negative impact for models that are designed to look for general patterns such as TopDBNet, causing a significant performance loss when dealing with whole image cases as compared to when it was trained using whole images. But overall there is a positive trend in the ability of all the models in dealing with occluded cases.

6.2 Limitations and Future Works

In this thesis, we explored a simple implementation of an online open-world person ReID system that was met with very specific constraints. Because of this, we had to opt for lightweight models that we're older and not very accurate when compared to the SOTA approaches. There is room for further studies to improve on replacing the models and approaches that we use in order to develop faster and more accurate real-time ReID systems.

Furthermore, further studies can also be done to explore the use of better preprocessing and post-processing techniques to increase the robustness of the ReID system. As one of the main challenges with developing a system such as this is the very small margin of error, thus requiring research on methods that can ensure that the system is still able to make guesses even when the system itself lacks data or makes a mistake.

We also admit that we weren't able to fully explore the limits of the impact of occluded data in the training data of a model because to our best knowledge, no such dataset exists. This creates possible future directions as we have established the importance of occluded samples in addressing such cases in real-life ReID applications. New network structures can also be proposed to deal with occluded data and possibly address those cases without losing accuracy on the whole case. Additionally, with what we observed in Chapter 5, further study on the effects of different network structures and their training process could be explored as well. This would serve to provide insights on our hypothesis that the model's design plays a part in how these models interpret training data and the results of the training. Such directions are promising in the field of creating faster and more accurate ReID models for real-life applications.

References

- [1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [2] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” 2016.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
- [4] D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning deep context-aware features over body and latent parts for person re-identification,” 2017.
- [5] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Computer Vision, IEEE International Conference on*, 2015.
- [6] Z. Zhang, J. Wu, X. Zhang, and C. Zhang, “Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project,” 2017.
- [7] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *CVPR*, 2014.
- [8] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” 2021.
- [9] Z. Zheng, L. Zheng, and Y. Yang, “A discriminatively learned cnn embedding for person reidentification,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, p. 13, 2018.
- [10] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

- [11] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, “Omni-scale feature learning for person re-identification,” in *ICCV*, 2019.
- [12] —, “Learning generalisable omni-scale representations for person re-identification,” 2021.
- [13] B. Xie, X. Wu, S. Zhang, S. Zhao, and M. Li, “Learning diverse features with part-level resolution for person re-identification,” 2020.
- [14] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, “Abd-net: Attentive but diverse person re-identification,” 2019.
- [15] F. Herzog, X. Ji, T. Teepe, S. Hörmann, J. Gilg, and G. Rigoll, “Lightweight multi-branch network for person re-identification,” 2021.
- [16] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [17] R. Quispe and H. Pedrini, “Top-db-net: Top dropblock for activation enhancement in person re-identification,” *25th International Conference on Pattern Recognition*, 2020.
- [18] J. Chen, Y. Wang, and R. Wu, “Person re-identification by distance metric learning to discrete hashing,” in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 789–793.
- [19] W. Fang, H.-M. Hu, Z. Hu, S. Liao, and B. Li, “Perceptual hash-based feature description for person re-identification,” *Neurocomputing*, vol. 272, 07 2017.
- [20] G. Wang, S. Gong, J. Cheng, and Z. Hou, “Faster person re-identification.” In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [21] M. Baharani, S. Mohan, and H. Tabkhi, “Real-time person re-identification at the edge: A mixed precision approach,” *CoRR*, vol. abs/1908.07842, 2019. [Online]. Available: <http://arxiv.org/abs/1908.07842>
- [22] Q. Leng, M. Ye, and Q. Tian, “A survey of open-world person re-identification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1092–1108, 2020.
- [23] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, “Partial person re-identification,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4678–4686.
- [24] L. He, J. Liang, H. Li, and Z. Sun, “Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

- [25] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, “Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification,” in *IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [26] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, “Pose-guided feature alignment for occluded person re-identification,” in *ICCV*, 2019.
- [27] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, “High-order information matters: Learning relation and topology for occluded person re-identification.” In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [28] J. Zhuo, Z. Chen, J. Lai, and G. Wang, “Occluded person re-identification,” 2018.
- [29] J. W. Chin, K. Long Wong, T. T. Chan, K. Suhartono, and R. H. So, “An infrared thermography model enabling remote body temperature screening up to 10 meters,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 3870–3876.
- [30] 2021. [Online]. Available: https://docs.opencv.org/3.4.15/d4/d1b/tutorial_histogram_equalization.html
- [31] 2021. [Online]. Available: <https://pytorch.org/vision/stable/models.html>
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1106–1114. [Online]. Available: <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>
- [33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” 2017, cite arxiv:1704.04861. [Online]. Available: <http://arxiv.org/abs/1704.04861>

Appendix A

Comparison of Models

In this appendix, we will detail the technical details of the main models that were studied in this thesis in Chapter 5. The intent is to assist with any further research in the direction of understanding the effect of the design of the models themselves to their performance under occluded cases.

In the models that we study, we separate them into 3 main categories according to the focus on the learned features in their design. The 3 categories are:

- **Global Feature Representation:** models that are designed with a focus on this usually just take in the image as a whole without any specific attention mechanisms. Trusting that the model itself will be able to discover significant image features from the whole image.
- **Local Feature Representation:** in contrast to the global feature representation, the local feature representation usually has strategies that specifically target certain local regions (e.g. body parts). This is usually combined with local features to help the model have more sources of information to compare identities with.
- **Auxiliary Feature Representation:** models with this focus try to add additional information on top of the base training data to help it learn extra information. This added information doesn't naturally exist in the base training dataset and is usually added as an extra information source.

Knowing these categories, we divided the models according to the specification of the models and what they are focused on in the following table.

Model Name	Focus on Features	Key Network Feature
ResNet-18	Global	Direct image information
Bag Of Tricks	Global	Training process refinements
PLR-OSNet	Local	Global and local feature branch
TopDBNet	Auxiliary	Scene foreground emphasis
LightMBN	Local	Multi branch information sources

The breakdown of the layers of each network are as follows:

- **ResNet-18:** 4 convolutional layers. 1 Block/Layer. Each block contains 2 convolution, 2 batch normalization, and a ReLU layer.
- **Bag Of Tricks:** 4 convolutional layers. 3, 4, 6, 3 blocks in each layer. Each block contains 3 convolution, 3 batch normalization, and a ReLU layer.

- PLR-OSNet: 5 convolutional layers. Introduces an OSBlock, which contains 4 individual layers with (1, 2, 3, 4) blocks each. All blocks for this network contain 1-2 convolution, 1 batch normalization, 1 ReLU layer. The breakdown of the 5 convolutional layers:
 - 1 block
 - 2 blocks + 1 OSBlock
 - 1 block + 2 OSBlock
 - 2 branches of 1 block + 2 OSBlock
 - 2 branches of 1 block
- TopDBNet: 2 convolutional layers. Blocks for this network contains 3 convolution, 3 batch normalization, and a ReLU layer. The breakdown of the 2 layers:
 - 2 blocks
 - 4 layers of 3, 4, 6, 3 blocks
- LightMBN: divided into 1 backbone layer and 3 branches. Each block contains 1-2 convolution, 1 batch normalization, and a ReLU layer. The breakdown of the layers:
 - Backbone: 2 blocks + 3 OSBlock
 - Global branch: 2 layers.
 - * 1 block + 1 OSBlock
 - * 1 block + 2 OSBlock
 - Partial branch: 2 layers.
 - * 1 block + 1 OSBlock
 - * 2 blocks + 1 OSBlock
 - Channel branch: 2 layers.
 - * 2 blocks + 1 OSBlock
 - * 3 blocks + 2 OSBlock

Each of the network deploys their own method of pooling and averaging the results from the different convolutions that they have. All of them tries to batch things together into one final feature representation form to use for comparison between identities.