Universidade Estadual de Campinas
Instituto de Computação

Edgar Rodolfo Quispe Condori

Vehicle and Person Re-Identification: Methods and Applications

Reidentificação de Veículos e Pessoas: Métodos e Aplicações

CAMPINAS
2023

Edgar Rodolfo Quispe Condori

Vehicle and Person Re-Identification: Methods and Applications

Reidentificação de Veículos e Pessoas: Métodos e Aplicações

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientador: Prof. Dr. Hélio Pedrini

Este exemplar corresponde à versão final da Tese defendida por Edgar Rodolfo Quispe Condori e orientada pelo Prof. Dr. Hélio Pedrini.

CAMPINAS
2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Q48v

Quispe Condori, Edgar Rodolfo, 1994-
Vehicle and person re-identification : methods and applications / Edgar Rodolfo Quispe Condori. – Campinas, SP : [s.n.], 2023.

Orientador: Hélio Pedrini.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Redes neurais convolucionais. 2. Visão por computador. 3. Reconhecimento de padrões. 4. Aprendizado de máquina. 5. Redes neurais (Computação). I. Pedrini, Hélio, 1963-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações Complementares

Título em outro idioma: Reidentificação de veículos e pessoas : métodos e aplicações
Palavras-chave em inglês:
Convolutional neural networks
Computer vision
Pattern recognition
Machine learning
Neural networks (Computer science)
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Hélio Pedrini [Orientador]
Ronaldo Cristiano Prati
André Santanchè
Carlos Antônio Caetano Júnior
Alexandre Mello Ferreira
Data de defesa: 27-02-2023
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-1661-3720
- Currículo Lattes do autor: http://lattes.cnpq.br/6931027168436055

Universidade Estadual de Campinas
Instituto de Computação

Edgar Rodolfo Quispe Condori

Vehicle and Person Re-Identification: Methods and Applications

Reidentificação de Veículos e Pessoas: Métodos e Aplicações

Banca Examinadora:

- Prof. Dr. Hélio Pedrini
  IC/UNICAMP

- Prof. Dr. Ronaldo Cristiano Prati
  CMCC/UFABC

- Dr. Carlos Antônio Caetano Júnior
  Samsung

- Dr. Alexandre Mello Ferreira
  FUNCAMP

- Prof. Dr. André Santanchè
  IC/UNICAMP

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 27 de fevereiro de 2023

What matters in life is not
What happens to you but
What you remember
And how you remember it.

(Gabriel Garcia Marquez)

Acknowledgements

To my late grandparents, Toribia and Gregorio. I always have both of you in my thoughts to remind me of my farming roots and how astounding life can be.

I would like to start by thanking my advisor Helio Pedrini for supporting me and allowing me to work at Microsoft while I finished my PhD for the last three years. Also for all those Sunday morning and Friday evening meetings.

Thanks to Rosa Enciso and Yumao Lu because they lit the spark that working and finishing my PhD at the same time was a possibility.

Thanks to Gord Lueck, a supportive manager and good friend for helping me balance work and research in the last three years. Also thanks to my collaborators Cuiling Lan and Zhizheng Zhang at Microsoft Research Asia, for all their countless hours of discussions and feedback. Thanks to all my teammates at Microsoft who unknowingly inspired me to improve as a scientist.

Thanks to my family members, who even though they are far away, have always made me feel that I have a place to come back to when things are complicated.

Thanks to my friends in Peru, Brazil and United States, because they helped me keep my sanity and balance during the pandemic and enjoy my time beyond work.

Thanks to everyone from the Institute of Computing at University of Campinas (UNICAMP), because they made my life easier by helping me focus on my research rather than on bureaucratic steps. A special thanks to Shirlei, part of the security staff, for the deep and fun conversation on Sunday afternoons.

Finally, I would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 - and the Brazilian Education System that gave me international opportunities, funded my research, and gave me free quality education during my Master's and Doctorate degrees.

Resumo

A reidentificação (ReID) é um problema da área de visão computacional que visa combinar instâncias de entidades (por exemplo, pessoas, veículos ou bagagens) através de um sistema de câmeras que não se sobrepõem. Vários fatores tornam a tarefa desafiadora, tais como oclusões, condições de iluminação, configurações de câmera, diferentes pontos de vista e fundos complexos das cenas. Diferentes domínios de aplicação podem se beneficiar do problema de ReID, por exemplo, vigilância e segurança, rastreamento, ciência forense e robótica. Nesta tese, investigamos esta tarefa em um esquema amplo e abrangente. Como a pesquisa no ReID evoluiu para um cenário mais real, nossa pesquisa também segue esta tendência. Iniciamos com a proposição de um método supervisionado para ReID de pessoas que melhora a representação de atributos de uma rede neural, aprendendo informações discriminativas de regiões de baixa ativação. Em seguida, avançamos para um cenário com maior quantidade de identificadores (IDs) e desenvolvemos um método que alavanca eficientemente os rótulos de atributos para ReID de veículos. Este método destila informações de atributos específicos de tarefas em vez de seguir a literatura que utiliza todas as informações de atributos. Finalmente, aplicamos o ReID à outra tarefa, chamada Rastreamento de Múltiplos Objetos. Investigamos um problema menos explorado na literatura e mostramos que o uso adaptativo do atributo ReID em objetos altamente ocluídos durante o treinamento leva a um melhor desempenho. Avaliamos nossos três métodos propostos em conjuntos de dados amplamente utilizados e mostramos que os resultados são competitivos.

Abstract

Re-Identification (ReID) is a problem in the field of computer vision that aims to match instances of entities (for instance, people, vehicles, or luggage) across a system of non-overlapping cameras. Several factors make the task challenging, such as occlusions, lighting conditions, camera settings, different viewpoints, and complex scene backgrounds. Different application domains can benefit from the ReID problem, for example, surveillance and security, tracking, forensic science, and robotics. In this thesis, we investigate this task in a broad and comprehensive scheme. Since research in ReID has evolved into a more real-world scenario, our research also follows this trend. We start by proposing a supervised method for person ReID that enhances the embedding feature representation of a neural network by learning discriminative information from low activated regions. We then move to a scenario with a larger amount of identifiers (IDs) and develop a method that efficiently leverages attribute labels for vehicle ReID. This method distills task-specific attribute information rather than following the literature that uses all attribute information. Finally, we apply ReID to another task, called Multi-Object Tracking. We investigate a less explored problem in the literature and show that adaptive use of the ReID feature on highly occluded objects during training leads to better performance. We evaluate our three proposed methods on widely used benchmarks and show that the results are competitive.

List of Figures

List of Tables

List of Abbreviations and Acronyms

AC	Amelioration Constraint
AGNet	Attribute-guided Feature Learning Network
ANet	Attribute Network
AReID	Adaptive Re-Identification
AUC	Area Under the Curve
BDB	Batch DropBlock Network
CAM	Class Activation Mapping
CAMA	Class Activation Map Augmentation
CMC	Cumulative Matching Curve
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
DenseNet	Dense Convolutional Network
DF-CVCT	Camera Views, Vehicle Types and Colors Network
DFN	Detection False Negatives
DFP	Detection False Positives
DLA	Deep Layer Aggregation
DNN	Deep Neural Network
DP	Detection Precision
DR	Detection Recall
DTP	Detection True Positives
ECN	Exemplar Memory Network
F-RCNN	Faster Region-Based Convolutional Neural Network
FC	Fully Connected
FN	False Negatives
FP	False Positives
FPN	Feature Pyramid Network
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GN	Graph Network
GPU	Graphics Processing Unit

Grad-CAM	Gradient-weighted Class Activation Mapping
HA-CNN	Harmonious Attention Network
HPGN	Hybrid Pyramidal Graph Network
IANet	Integration-and-Aggregation Network
ID	Identity
IDF1	F1 for Identity matching
IDFN	False Negatives of Identity matching
IDFP	False Positives of Identity matching
IDP	Precision of Identity matching
IDR	Recall of Identity matching
IDSW	Identity Switches
IDTP	True Positive of Identity matching
IOU	Intersection-over-Union
JM	Joint Module
k-Reciprocal	Re-Ranking based on k-reciprocal Encoding
LIV	Laboratory of Visual Informatics
LSTM	Long Short-Term Memory
MADVR	Multi-Attribute Driven Network
mAP	Mean Average Precision
MEB-Net	Multiple Expert Brainstorming Network
ML	Machine Learning
MLP	Multi-Layer Perceptron
MOT	Multi-Object Tracking
MOTA	Multi-Object Tracking Accuracy
MS-COCO	Microsoft Common Objects in COntext dataset
NLP	Natural Language Processing
NN	Neural Network
OSNet	Omni-Scale Network
P-ReID	Person Re-Identification
PAMAL	Partial Attention and Multi-Attribute Network

PAMTRI	Pose-Aware Multi-Task Network
PCRNet	Parsing-guided Cross-part Reasoning Network
PGAN	Part-Guided Attention Network
PL	Pseudo-Labels
PRND	Part-Regularized Near-Duplicate Network
PVEN	Parsing-based View-aware Network
ReID	Re-Identification
ReLU	Rectified Linear Unit
RGB	Red-Green-Blue
RK	Re-Ranking
RNN	Recurrent Neural Network
SAL	Self-Supervised Agent Learning
SAN	Stripe-based and Attribute-Aware Network
SAVER	Self-Supervised Attention Network
SONA	Second-Order Non-Local Attention Network
SP	Semantic Parsing
SPHM	Self-Paced Contrastive Learning Network
StRDAN	Synthetic-to-Real Domain Adaptation Network
TDM	To Drop Mask
Top-BDB-Net	Top DropBlock Network
TS	Teacher Student Network
UDA	Unsupervised Domain Adaptation
UMTS	Uncertainty-Aware Multi-Shot Network
UNICAMP	University of Campinas
UNRN	Uncertainty-guided Noise Resilient Network
V-ReID	Vehicle Re-Identification
VAD	Vanilla-Attribute Design
VAN	VAD-based Network
Vanet	Viewpoint Aware Network
VGG	Visual Geometry Group
VKD	View Knowledge Distillation
VPD	Vehicle Part Detection
XBM	Cross-Batch Memory for Embedding Learning

Contents

Chapter 1

Introduction

As cities grow and more security cameras are available in public spaces, there is an increasing necessity for intelligent surveillance and monitoring systems. An important capability of these systems is to match entities across non-overlapping views, also known as Re-Identification (ReID).

ReID has various applications in real-world problems. For instance, during the investigations of the 2013 Boston Marathon Bombings, police officers went through a great deal of tedious work to match and track the entities of the attackers across cameras in restaurants, grocery stores, and parking lots (see Figure 1.1). Similarly, law enforcement officers have gone through a slow and tedious process to follow hit-and-run incidents, and the number of fatalities in this type of accidents has increased 26% between 2020 and 2022 in the United States, which is an alarming statistic. Having proper ReID systems in place would make these processes more efficient. Moreover, in the case of hit-and-run incidents, a tracking system would help to follow suspects in crowded highways. Other ReID applications include robotics, forensics, retail and healthcare [87].

In this work, we investigate the ReID problem in several scenarios. Moreover, we explore how to apply ReID to a closely related task, known as Multi-Object Tracking. Different approaches were proposed and evaluated on known datasets in these tasks.

We start by exploring the first proposed scenario for the ReId problem: Person Re-Identification (P-ReID). Specifically, we explore how to generate a more reliable feature representation. Then, we study a scenario that includes a larger amount of entities, that is, Vehicle Re-Identification (V-ReID), we leverage extra labels to train a more generic and invariant model. Finally, we explore an overlooked problem when training models for Multi-Object Tracking (MOT) using Re-Identification as subtask. In Particular, we show that Re-Id information should be used adaptively.

1.1 Problem Statement

Re-Identification (ReID) is a challenging computer vision task defined as a retrieval task: given a query image or video, and a gallery of images or videos, ReID aims to retrieve all the instances in the gallery that have the identity (ID) as a query (See Figure 1.2). In this work, we focus on the scenario based on images. The setup of ReID is particular since

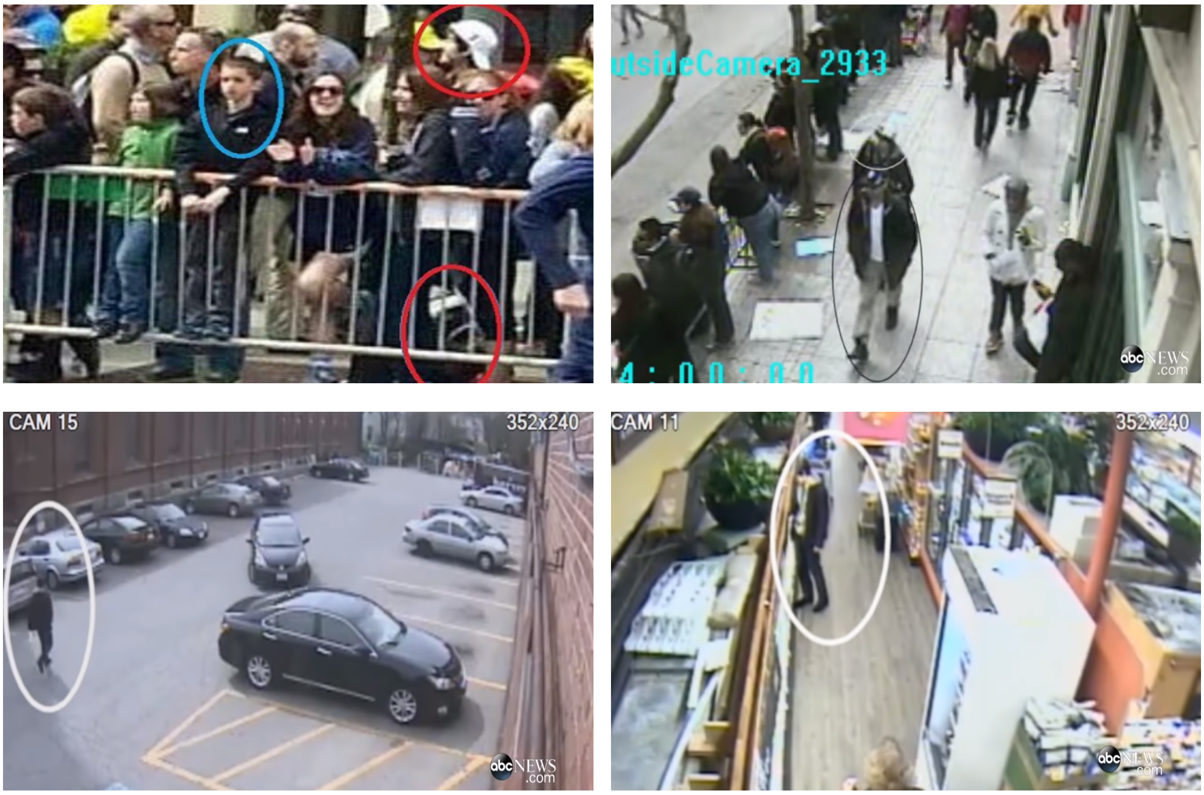


Figure 1.1: Images of the 2013 Boston Marathon Bombings. Attackers are highlighted in red, black, and white colors.

it is an open-set scenario, that is, the classes/IDs used during training are not the same during testing.

ReID derived from Multi-Person Tracking, so its initial version was mostly focused on people, also known as Person ReID. However, there has been an increasing availability of methods for new scenarios (for instance, Vehicle ReID and Luggage ReID [80, 116]). In this work, we focus on Person ReID (P-ReID) and Vehicle ReID (V-ReID).

Person ReID considers that we use RGB images. The images of the query and gallery are already from the bounding boxes of people (See Figure 1.2), whose images are typically captured with security cameras, so clues such as the faces themselves are not usable. In this setup, the ReID is mainly based on the appearance of the people (for instance, pants color, jacket type). Examples of these images are illustrated in Figure 3.1. There are some variations of Person ReID that include infrared images, portrait-to-image [39, 40, 41, 100, 105]. However, we focus only on the *classic setup* in this work.

Vehicle ReID is analogous to the Person ReID problem, where the images are RGB of the bounding boxes, captured with security cameras, and the license plates are not a clue that we can use. However, the datasets for Vehicle ReID made available attribute labels such as vehicle model, type and color. Examples of these images are illustrated in Figure 3.2.

Multi-Person Tracking is a specific scenario of Multi-Object Tracking (MOT). MOT is an important computer vision task that aims to identify and track objects of a target

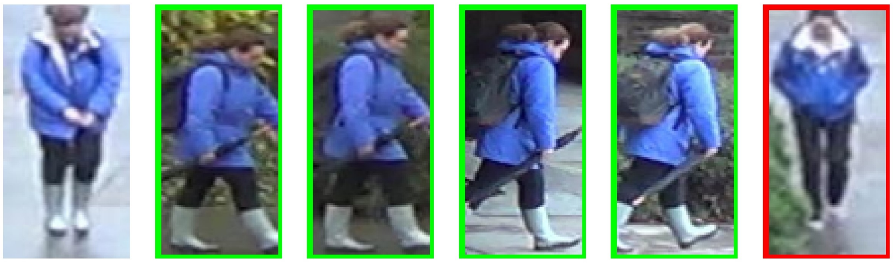


Figure 1.2: Example of bounding boxes for query (first image) and gallery (last 4 images). Gallery images with the same ID as the query are border in green, and red otherwise.

class (for instance, Vehicles and People).

There are multiple research lines in MOT, including tracking-by-detection [55, 74, 111, 113, 114], tracking-by-regression [2, 4, 52, 91] and, more recently, tracking-by-attention [67]. In this work, we focus on tracking-by-detection since it has been widely investigated. This paradigm has two steps: (i) object detection and (ii) matching affinity. We focus on leveraging ReID to further improve them. More specifically, we focus on MOT applied to people because it is closer to the original ReID task definition.

ReID and MOT are closely related, and actually it is possible to argue that matching affinity could be solved using ReID. However, MOT can leverage the hypothesis of constancy to use motion and intersection-over-union (IOU) for matching. Moreover, in this work, ReID is based on images and MOT is based on videos.

1.2 Challenges

We focus on various ReID scenarios, which include P-ReID and V-ReID using supervised learning. Due to that, the challenges can be separated into data-related and fashion-related.

Data-related challenges for both P-ReID and V-ReID include occlusions, different viewpoints, illumination conditions, background clutter, and camera settings. These problems are especially complex because images are usually captured over large periods of time (for instance, months) in unconstrained scenarios, such as public markets, university campuses, and city streets. In the case of P-ReID, there is an extra challenge generated by non-rigid deformation of human bodies. In the case of V-ReID, there is an extra challenge for the situations where two vehicles with different IDs share the same model/manufacturer.

Fashion-related challenges are associated with the type of approach used. In the context of supervised methods, the main challenge is feature representation: how to train a model that can represent the ID of the image being invariant to the previous data-related challenges? Previous works explore different ideas to solve this problem, such as semantic parsing, view invariant models using camera information, attributes, attention, adversarial learning, among others.

In addition, we also explore MOT. Similarly to ReID, the challenges can be separated into data-related and fashion-related. Data-related challenges in MOT are basically the same as in ReID; however, the main challenges are the simultaneous occlusions and interactions between objects.

In this work, we follow the tracking-by-detection fashion. Thus, the fashion-related challenges are associated with two of the steps involved (for instance, detection and matching). For the detection step, the main problem is related to occlusions, different object sizes, and intra-class variance. For the matching step, the main challenge is to recover from the missing detections and the object feature representation that is highly related to ReID.

1.3 Objectives

The main purpose of this thesis is to *study and propose novel methods to tackle the challenges of re-identification and its applications.* To accomplish this goal, we stated the following objectives:

O1: Implement frameworks for experimentation, validation, evaluation and analysis of re-identification and multi-object tracking approaches.

O2: Propose a method that explores areas of images with low activations to create an enhanced feature representation.

O3: Develop an approach that leverages attribute labels to create a generic representation.

O4: Propose a method that leverages re-identification for multi-object tracking.

Therefore, this work is guided to answer the following research questions:

Q1: Is it possible to encode rich information from areas of the image that previous methods consider less relevant for ReID?

Q2: Is all the generic attribute information relevant for ReID?

Q3: How much are occlusions overlooked when applying ReID features to MOT?

1.4 Contributions

The main contributions of our work are listed as follows:

- **ReID Representation without extra labels**: Not all parts of the input images are equally relevant to the final feature representation, previous works would focus on pushing the network to learn task-relevant regions while ignoring low informative regions. We propose a method that pushes the network to learn to represent low-informative regions with more discriminative features. In doing so, the feature quality of task-relevant regions is further improved, creating a better overall final feature representation.

- **ReID Representation with extra labels**: ReID using extra labels (for instance, object features such as color and type) has been previously explored using a simple and ineffective design. We propose a novel method that increases the interaction between the extra labels and image features to create a more robust final representation. Furthermore, we propose a generic module that creates a compensated feature representation that is by definition always better than non-compensated representation.

- **ReID applied to MOT**: ReID applied to MOT has ignored an important problem, where occlusions are too difficult to handle by ReID. Because of this, current MOT designs tend to confuse the features of different ID objects when objects of the target class occlude each other. We propose a novel adaptive weight method that will use ReID features only in the cases where it can be really beneficial.

1.5 Publications

During the development of this research work, we published Top-DB-Net [79] and AttributeNet [80], which explore research questions Q1 and Q2, respectively. Furthermore, we pre-printed ReID guided MOT [81], which explores research question Q3:

1. R. Quispe, H. Pedrini. "Top-BDB-Net: Top Batch DropBlock for Activation Enhancement in Person Re-Identification". *25th International Conference on Pattern Recognition* (ICPR), pp. 2980-2987. Milan, Italy, January 2021.

2. R. Quispe, C. Lan, W. Zeng, H. Pedrini. "AttributeNet: Attribute Enhanced Vehicle Re-Identification". *Neurocomputing*, vol. 465, pp. 84-92, November 2021.

1.6 Text Organization

This text is organized as follows. In Chapter 2, we review the literature focusing on related methods for both ReID and MOT. In Chapter 3, we introduce the metrics used in the result evaluation, data sets and hardware/software used in the experiments. In Chapters 4, 5 and 6, we present and discuss our developed methods. Finally, we describe the concluding remarks of this thesis and some directions for future work in Chapter 7.

Chapter 2

Background

The term ReID was first stated by Zajdel et al. [109] as a variation to the people tracking problem [95]. In this chapter, we introduce a background for ReID and MOT, which includes fundamentals and literature review of relevant related work.

2.1 Fundamentals

This section briefly describes some relevant concepts related to the Multi-Object Tracking and Re-Identification problems, which aim to help in the understanding of the techniques developed in this work.

2.1.1 Deep Neural Networks

Machine Learning (ML) [1, 3, 130] has evolved and moved to the use of Deep Neural Networks (DNNs) in the last ten years. Previously, ML methods focused on designing features that could accurately represent the input to the intended task. Later, DNNs created a new paradigm where these features are automatically learned from the data and research has focused on designing and guiding DNN to learn task-relevant features.

In this work, DNNs are used to tackle the challenging ReID and MOT. The type of DNNs we most used in this work are Convolutional Neural Networks (CNNs). CNNs became really popular for computer vision tasks after the outstanding results they achieved in the ImageNet Challenge [83]. Over the years, the use of Transformers [94] has become more popular. In this work, we do not use Transformers directly in our methods, but we compare our results with them.

In our first two approaches, we use ResNet [31]. ResNet was designed to efficiently train deep models by adding a skip connection between layers, which allows information to flow better during the learning process. In our third work, we use networks based on ResNet [31], but with extra modifications such as Deep Layer Aggregation (DLA) [129] and U-Net [10].

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a specific type of Deep Neural Network. The main difference between this type of network is the use of convolutional kernels that become the correlation operation [25]. These kernels learn to extract patterns from small parts (e.g., active fields) from the input image. Stacking multiple layers of these convolutional layers allows CNNs to learn complex patterns from input images. For instance Figure 2.1 shows learned patterns at different depth layers of a CNN.

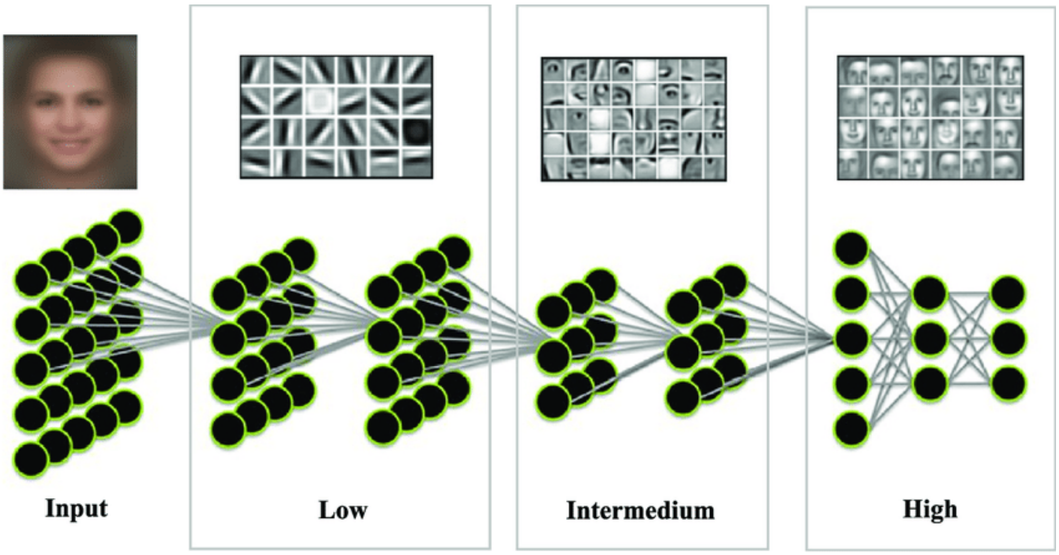


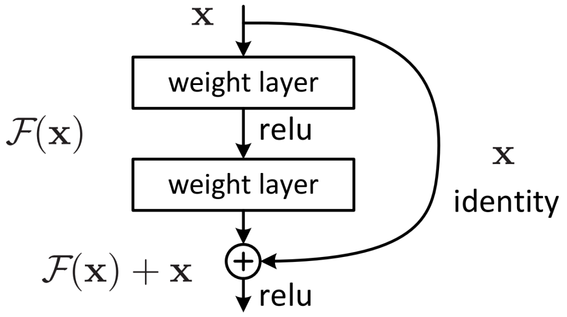
Figure 2.1: Examples of CNN learned patterns. At the initial layers, the network learns to recognize lines in different orientations and is able to recognize human faces in the deepest layers [54].

ResNet

ResNet is a specific type of CNN [31]. It was designed to alleviate the vanishing gradient problem. In order to do so, a skip connection is added to each layer, such that the gradient from deeper layers has a direct backpropagation path to the shallower layers. This CNN is important for our work since we used it as the backbone for all proposed methods. Figure 2.2 shows the implementation of the residual block, which is the core part of a ResNet.

2.1.2 Attention

The concept of attention [29, 72] in DNNs refers to the ability of networks to learn and focus to extract more information from certain areas of input images. In this work, we use this concept to improve the overall feature representation and to analyze the effectiveness of our methods. This concept is closely related to Transformers [94]. These types of networks are generic feature extractors that were initially used in Natural Language Processing (NLP) [9, 17, 70], but were later also adopted in computer vision. The main advantage of attention-based methods is that they assume no prior from the data, are versatile and can be used in multi-model learning. However, they also require more time and data to train.



x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2.2: Implementation of a residual block. The identity connection allows to alleviate the vanishing gradient problem [31].

In our first approach, we explicitly use the attention clue to push our network to learn stronger features. In our second approach, we show that our method improves the attention that the network learns to represent attributes.

2.1.3 Re-Ranking

Re-Ranking has been widely studied in the context of image retrieval [123]. This process is described as a refinement step after an initial set of documents (e.g., bounding boxes for ReID) are retrieved. Figure 2.3 shows an example of the Re-Ranking process where the distance between retrieved candidates is further used for consistency matching and refinement.

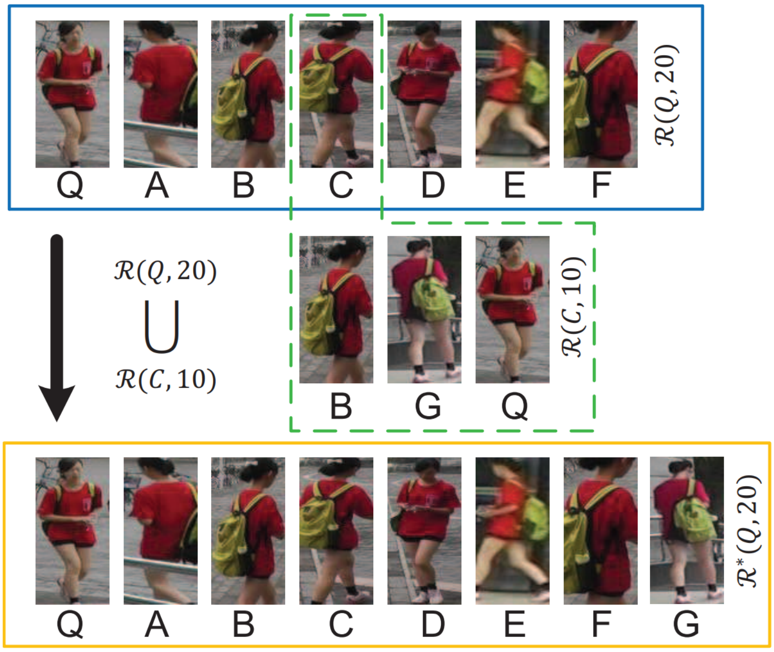


Figure 2.3: Example of ReID process. First row shows the initial set of retrieved bounding boxes for query Q, second row shows the list of neighbors (e.g., list of possible matches) used to refine results, and third row shows the re-ranked results [123].

2.2 ReID

Although ReID has its origins in tracking, unlike tracking algorithms, ReID does not depend on the hypotheses of constancy. Thus, it is a more complex problem due to considerable variations in biometric profile, position, appearance and point of view [95].

In this section, we focus on approaches that are strongly related to our proposed methods [79, 80], which includes P-ReID and V-ReID. For each scenario, we include a generic analysis and a table that summarizes the main ideas of the analyzed methods.

P-ReID

Many early works in the ReID context considered it as a classification task. This is mainly due to the fact that most of the datasets [27] available at that time had only a few instances of each person. Because of this, various methods based on handcrafted features [8, 13, 34, 35, 50] were initially proposed.

With the popularization of deep learning and ReID, many datasets with larger amounts of instances per person in real-world scenarios have been made available [52, 82, 119, 121, 123] and deep networks-based methods have become the standard [126, 128].

This has brought two side effects: (i) the most popular datasets already include a predefined training and testing splits – which helps with validation protocols and comparison among methods – and (ii) ReID turned into a retrieval problem – thus, measures such as Cumulated Matching Characteristics (CMC) and Mean Average Precision (mAP) are widely used.

Several proposed methods for ReID use specific prior related to a person's nature, such as pose and body parts [7, 47, 51, 78, 120]. However, labels such as segmented semantic regions and body skeleton that are required for these types of methods are not available in current ReID datasets. Thus, they usually leverage datasets proposed for other tasks captured in different domains, which introduces noise during training and makes the learning process more complicated.

On the other hand, there are methods [14, 37, 53, 104] that learn to encode rich information directly from the input images without relying on other types of signals. Most methods in this category use the concept of attention in their pipeline. Thus, their approaches expect the networks to learn to focus on discriminative regions and encode those parts. However, assuming that the availability of consistently discriminative regions may introduce errors, since occlusions are a major problem in the context of ReID due to drastic view changes.

To further improve ReID performance, the literature has proposed re-ranking [84, 123] and augmentation [122, 124] approaches. The former methods can improve ReID results by a huge margin, which makes it unfair to compare pipelines that use them with pipelines that do not use them. Therefore, since several state-of-the-art methods report results with and without re-ranking, our comparison with them is made separately for these two scenarios.

Table 2.1 shows an overview of related methods for P-ReID.

Table 2.1: Overview of relevant methods for P-ReID.

Method	**Key Ideas**
BDB [14]	Erases random regions of feature tensor to enhance activations.
IANet [37]	Models interdependencies between spatial features and combines correlated features between body parts, which increases the feature representation capability of the network.
HA-CNN [53]	Focuses on improving methods based on bounding methods by learning a combination of soft pixel attention and hard regional attention along with simultaneous optimization of feature representations.
SONA [104]	Models local and non-local relationships via second-order feature statistics.
k-Reciprocal [123]	Re-Ranking method based on hypothesis: if a gallery image is similar to the probe in the k-reciprocal nearest neighbors, it is more likely to be a true match.

V-ReID

For vehicle ReID, many approaches explore Generative Adversarial Networks (GANs) [45], Graph Networks (GNs) [63, 86], Semantic Parsing (SP) [68] and Vehicle Part Detection (VPD) [30, 112] to improve performance. Some of them tend to describe vehicle details [45] and local regions [30, 112].

PRND [30] and PGAN [112] detect predefined regions (for instance, back mirrors, light, wheels, among others) and describe them with deep features. Some works aim to handle drastic viewpoint changes [63, 68].

Some works explore attribute information [28, 58, 62, 76, 98] or combine attributes with other clues [49, 92, 117]. Most previous attribute-based works use attribute information to regularize the feature learning [49, 58, 62, 76, 92, 98, 117]. In general, they regress the attribute classes from the backbone features, along with the ReID supervision based on the backbone features. However, using separate heads for different tasks ignores the interaction between the two tasks, where the attribute branches should serve for better ReID.

Table 2.2 shows an overview of related methods for V-ReID.

2.3 MOT

MOT Methods follow different fashions, including tracking-by-detection [55, 74, 111, 113, 114], tracking-by-regression [2, 4, 52, 91] and more recently tracking-by-attention [67]. In this work, we focus on tracking-by-detection because its interaction with ReID has been widely studied.

Tracking-by-detection methods have two steps: detection and association. In the first step, each frame is processed independently to detect the bounding boxes of the target class objects. In the second step, the relationship of these bounding boxes is analyzed to assign

Table 2.2: Overview of relevant methods for V-ReID.

Method	**Key Ideas**
SAVER [45]	Modifies the input image with the vehicle details erased using a GAN. Then, this synthetic image is combined with the input image to create a new version with the details visually enhanced for ReID.
StRDAN [49]	Leverages synthetic and real data to improve the feature representation capability.
PCRNet [63]	Describes each vehicle view based on semantic parsing and also encodes the spatial relationship between them using GNs.
SAN [76]	Proposes a stripe based network that is combined with vehicle attribute learning. In this case, both stripe and attribute heads share the backbone.
PAMTRI [92]	Estimates vehicle viewpoint to extract key points, segments and heatmaps. These clues aim to create a view invariant feature. Moreover, augmentation based on synthetic images and vehicle attributes are used.
AGNet [98]	Uses vehicle attributes to define an attention mask to filter information from the backbone feature tensor.
DF-CVCT [117]	Uses a GAN to augment viewpoints and leverages vehicle attributes as gates to regulate the flow of information of different layers of the backbone in the final representation.

an ID to them. Most of the works tend to focus on the detection step [55, 74, 111, 113] and a few of them on the second step [114]. Our work also focuses on the first step. Therefore, in this review, we will focus on those methods. However, we will describe ByteTrack [114] because we use it in order to compare our method against the state of the art.

Table 2.3 shows a list of related methods for MOT.

2.4 Final Considerations

In this chapter, we introduced key concepts and methods related to our research work. In the next chapter, we introduce the validation protocols, including metrics and datasets that we used to validate our approaches.

Table 2.3: Overview of relevant methods for MOT.

Method	**Key Ideas**
SAVER [45]	Modifies the input image with the vehicle details erased using a GAN. Then, this synthetic image is combined with the input image to create a new version with the details visually enhanced for ReID.
JDE [101]	One of the first works that proposed to use ReID combined with Detection for MOT. This approach aims to work in real-time and is anchor-based.
FairMOT [113]	This is an evolution on top of JDE [101] as it follows an anchor-free design. This change pushed the performance and made this method the best at the time of its publication.
CSTrack [55]	This is another evolution on top of JDE [101], which focuses on the competition between detection and ReID tasks, presenting a disentangling method that separates the information learned by each head.
MOTR [111]	This is a transformer-based method that aims to unify detection and association into an end-to-end process. It includes a temporal aggregation module that aims to track the changes of each branch.
ByteTrack [114]	This method focuses on the association step and proposes a two-step process that uses previously ignored detections and associates them using a more simple feature representation.
QuasiDense [74]	This method introduces dense matching to associate candidate bounding boxes at each pixel. It combines this process with similarity learning (for instance, ReID).
TraDes [103]	This approach introduces tracking clues in the process of detection learning. It learns to infer offset by a cost volume, which improves segmentation and detection.

Chapter 3

Materials

In this chapter, we present the metrics and datasets used to validate our methods. In addition, we describe the hardware and software resources used during the experiments.

3.1 Metrics

This section describes the metrics employed to assess the achieved results in the problems investigated in this thesis.

3.1.1 ReID

ReID can be considered an information retrieval task where, for a given query image, we aim to retrieve a list of candidate images sorted by the distance to the query and we expect them to have the ID as the query. Because of this, the most common metrics to evaluate ReID methods are: Cumulative Matching Characteristic (CMC) curve and mean Average Precision (mAP).

- **CMC**: This metric represents the probability that a correct match to the query ID appears in the first k items of the candidate list. CMC ignores the amount of ground-truth matches that are in the gallery because only the first match is counted. In the CMC curve, the abscissa axis represents the k-th rank, whereas the ordinate axis is defined as

  $$f(k) = \frac{in(k)}{\#queries} \tag{3.1}$$

  where $in(k)$ is the number of queries that have a relevant element inside the first k items recovered. To compare methods it is common to consider $k = 1$, denoted as rank-1/R1, and $k = 5$, denoted as rank-5/R5.

- **mAP**: Different to CMC, mAP metric considers all the retrieved items. It also considers the order in which they are retrieved. mAP is defined as

  $$\text{mAP} = \frac{\text{AP}}{\#queries} \tag{3.2}$$

where AP is defined as

$$\mathrm{AP} = \frac{\sum\limits_{k=1}^{n} P(k)\cdot \mathit{rel}(k)}{\#\mathit{relevant\ items}} \tag{3.3}$$

where n is the number of retrieved items, $\mathit{rel}(k)$ is equal to 1 if the k-th item is relevant for the query and 0 otherwise, and $P(k) = \frac{\sum\limits_{i=1}^{k} \mathit{rel}(i)}{k}$.

3.1.2 MOT

MOT algorithms can be measured using various types of metrics, where each of them focuses on different aspects of the algorithms. For instance, ID metrics focus on how many of the detected bounding boxes have a correctly assigned ID, while other metrics focus on how good the detected bounding boxes are. In this work, we used CLEAR metrics, Detection Metrics and ID Metrics since this thesis focuses on ReID.

- **CLEAR Metrics**: These metrics have become popular as they consider both the quality of the detected bounding boxes and the assigned IDs. We use on Multi-Object Tracking Accuracy (MOTA), which is expressed as:

  $$\mathrm{MOTA} = 1 - \frac{\mathrm{FN} + \mathrm{FP} + \mathrm{IDSW}}{\mathrm{GT}} \tag{3.4}$$

  where

  - FN: number of target bounding boxes that were not detected.
  - FP: number of bounding boxes detected that are not target.
  - IDSW: number of identity mismatches. This is, if a target x is assigned to track y and the last known assignment was $z \neq y$.
  - GT: number of ground truth bounding boxes.

- **Detection Metrics**: These metrics focus on the bounding box detected. We use Detection Precision (DP) and Detection Recall (DR), which are defined as:

  $$\mathrm{DP} = \frac{\mathrm{DTP}}{\mathrm{DTP} + \mathrm{DFP}} \tag{3.5}$$

  $$\mathrm{DR} = \frac{\mathrm{DTP}}{\mathrm{DTP} + \mathrm{DFN}} \tag{3.6}$$

  where

  - DTP: number of True Positive Detections.
  - DFP: number of False Positive Detections.
  - DFN: number of False Negative Detections.

- **ID Metrics**: This type of metrics compare the ground truth IDs of each bounding box against their predicted ID. We use IDF1, which is defined as:

$$\mathrm{IDF1} = \frac{2}{\frac{1}{\mathrm{IDP}} + \frac{1}{\mathrm{IDR}}} \tag{3.7}$$

where

$$\mathrm{IDP} = \frac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFP}} \tag{3.8}$$

$$\mathrm{IDR} = \frac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFN}} \tag{3.9}$$

  - IDTP: number of bounding boxes with ID correctly assigned.
  - IDFP: number of bounding boxes with false positive IDs assigned.
  - IDFN: number of bounding boxes with false negative IDs assigned.

3.2 Datasets

This section presents the main aspects related to the datasets used in our experiments.

3.2.1 ReID

In this work, we tackle two ReID scenarios: Person ReID and Vehicle ReID. We show an overview of the main characteristics of all the datasets in Table 3.1 and examples of images in Figures 3.1 and 3.2.

Table 3.1: Overview of characteristics of datasets used for ReID.

Dataset	**Release Time**	**# Images**	**# Cameras**	**# IDs**
CUHK03	2014	13,164	10	1,467
Market1501	2015	32,217	6	1,501
DukeMTMC-ReID	2017	36,441	8	1,812
Vehicle-ID	2016	221,763	2	26,267
VeRi776	2017	50,000	20	776
VeRi-Wild	2019	416,314	174	40,671

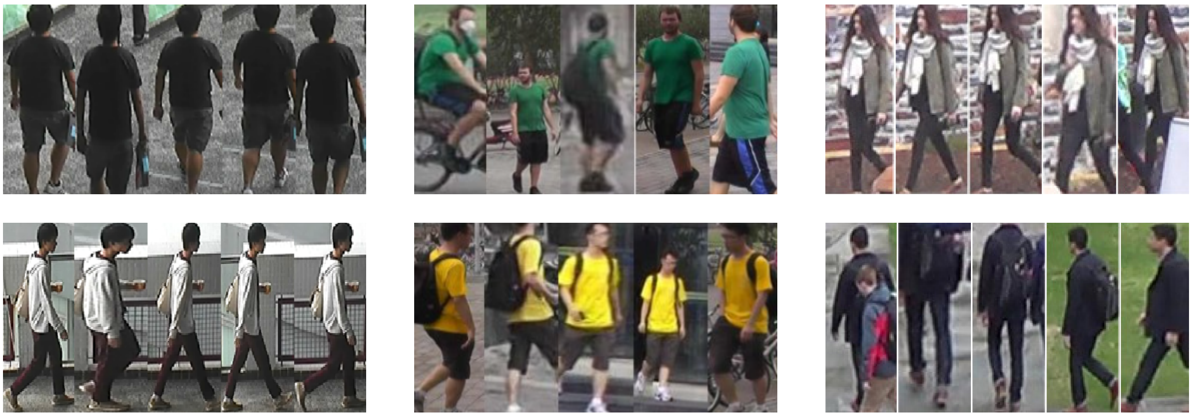
Person ReID

- **CUHK03** [52]: It was the first dataset large enough for ReID. The images are obtained from various months of recordings at the Chinese University of Hong Kong. Initially, the validation protocol was based on 20-fold validation, but later it was changed to fix a split that includes 767 IDs for training and 700 IDs for testing. It exhibits recurrently missing body parts, occlusions and misalignment. We tested its two versions: CUHK03 (D), that uses bounding boxes detected the Deformable

Part Model (DPM) [19], and CUHK03 (L), that includes manually labeled bounding boxes.

- **Market1501** [119]: It aims to simulate a more real-world scenario, where the images are captured in a market in front of a campus supermarket using five 12080 × 1080 HD cameras and one 720 × 576 SD camera. The bounding boxes are generated through the Deformable Part Model (DPM) [19], but their quality is in general worse than CUHK03(D). This dataset has been commonly used in the ReID literature to fix training/testing splits, containing 750 training IDs and 751 testing IDs.

- **DukeMTMC-ReID** [82, 121]: This is the largest dataset we used for Person ReID. It was originally designed for tracking and then adapted for ReID. It has hand-drawn bounding boxes with various backgrounds of outdoor scenes. The validation is based on 702 IDs for training and 702 IDs for testing.



CUHK03 Market1501 DukeMTMC-ReID

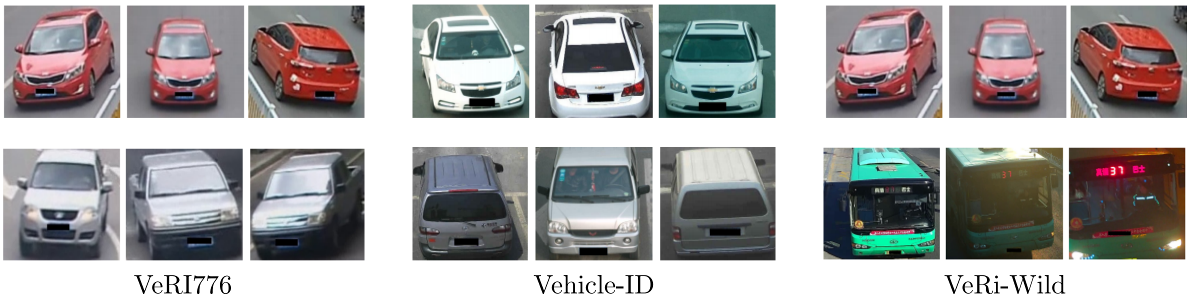
Figure 3.1: Examples of images for Person ReID datasets used in our work.

Vehicle ReID

- **VeRi776** [61]: It contains images captured from real-world unconstrained surveillance scenes and includes attribute labels for vehicle color and type. The images present different viewpoints, illumination, resolutions and occlusions. It includes information about license plates, but we do not use them. It considers 576 IDs for training and 200 IDs for testing.

- **Vehicle-ID** [58]: It includes images captured during daytime from multiple surveillance cameras distributed in a small city in China. It has images captured from either the front or back views. The training set contains 13,134 IDs and the testing set contains IDs 13,133 vehicles. The testing data is further divided into three sets with 200 (small), 1,600 (medium) and 2,400 (large) IDs. Some images in this dataset have attribute labels for vehicle color and type.

- **VeRi-Wild** [64]: This is the largest vehicle ReID dataset, where images were captured from a large urban district (more than 200 km^2 captured for a period of one month for 24 hours. Therefore, it is considered a realistic dataset, containing severe changes in background, illumination, viewpoint and occlusions. It also includes attribute labels for vehicle model, color and type. The training set has 40,671 IDs and the testing set has 10,000 IDs. The testing set is divided into three sets with 3,000 (small), 5,000 (medium) and 10,000 (large) IDs.



VeRI776 Vehicle-ID VeRi-Wild

Figure 3.2: Examples of images for Vehicle ReID datasets used in our work.

3.2.2 MOT

In this work, we focus on MOT using people as the target class. Therefore, we use the two most popular datasets [15, 69] in this setup. During training, we leverage pre-training in datasets such as MS-COCO [56] and Crowd Human [85], but they are not the target datasets for our analysis, so we do not describe them in detail in this section. Table 3.2 shows the main characteristics of MOT17 [69] and MOT20 [15]. Figure 3.3 shows some image examples.

Table 3.2: Overview of characteristics of datasets used for MOT.

Dataset	Release Time	# Tracks	# Boxes	Avg. Density
MOT17	2017	1,331	300,373	21.6
MOT20	2020	1,501	765,465	170.9

- **MOT17** [69]: This dataset was an evolution of MOT15 [48] as it focused on more realistic scenario. The ground truth were labeled from scratch using several detection algorithms with a more consistent protocol. In our work, we focus on the ground truth generated using a Faster R-CNN [24]. It has 7 sequences for training and 7 for testing. The testing set is closed, so during ablation we split the training set in half to analyze our proposed method. Some of its main challenges include different viewpoints, various weather and illumination conditions, and with/without camera motion. Extra labels include sitting people, vehicles and occluding objects.

- **MOT20** [15]: It was proposed to further challenge the MOT scenarios. It has 8 sequences with very crowded scenes with up to 246 pedestrians per frame. The sequences were recorded both outdoors and indoor, including day and night. This dataset has more than twice as many bounding boxes as MOT17 [69]

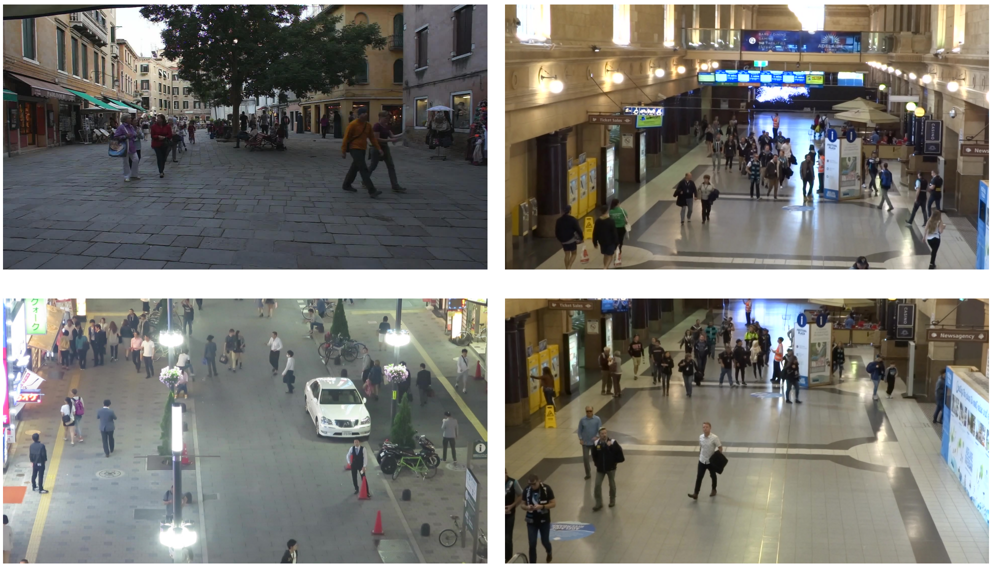


Figure 3.3: Examples of images for MOT used in our work.

3.3 Hardware and Software Resources

All experiments require intensive computational power, especially due to the large amount of images and size of the networks. Our main resources are graphics processing units (GPUs), which allow faster training of Deep Neural Networks.

The devices are available in the Laboratory of Visual Informatics (LIV) of the Institute of Computing at University of Campinas (UNICAMP), which is equipped with GeForce GTX 1080 Ti and TITAN V GPU cards with 11 and 12 GB of memory, respectively. In addition, we also used clusters from Microsoft Research, specifically nodes with 1, 2, 4 and 8 Tesla v100 GPUs (standard NDv2-series), each GPU with 16 GB of memory.

The main programming language is Python due to the availability of a large number of libraries for deep learning, numeric computation and image processing. Some libraries that can be highlighted are `NumPy`[1], `SciPy`[2], `scikit-learn`[3], `PyTorch`[4] and `Matplotlib`[5].

[1] https://www.numpy.org

[2] https://www.scipy.org

[3] https://scikit-learn.org

[4] https://www.pytorch.org

[5] https://matplotlib.org

3.4 Final Considerations

In this chapter, we introduced the validation protocol and tools used during the development of our research work. In the next chapter, we introduce our first method for ReID, which is specifically designed for person ReID and aims to create strong feature representation.

Chapter 4

Top-DB-Net: Learning Enhanced Features from Less Activated Regions

In this chapter, we introduce Top-DB-Net, a supervised method for Person Re-Identification (P-ReID) that focuses on improving the feature embedding learned by the DNN through the exploration of information from low activated regions.

4.1 Introduction

Numerous approaches have been proposed using person-related information, such as pose and body parts [7, 47, 51, 78, 120]. However, P-ReID datasets only provide ID labels. Thus, these methods rely on other datasets proposed for related tasks during the training. This dependency introduces further errors in predictions and motivates the creation of general methods that do not learn from outer information.

In this section, we introduce the Top DropBlock Network (Top-DB-Net) for the P-ReID problem. Top-DB-Net is designed to further push networks to focus on task-relevant regions and encode low informative regions with discriminative features.

Our method is based on three streams consisting of (i) a classic global stream as most of the state-of-the-art methods [7, 14, 47, 51, 65, 78, 120], (ii) a second stream drops[1] most activated horizontal stripes of feature tensors to enhance activation in task-discriminative regions and improve encoding of low informative regions, and (iii) a third stream regularizes the second stream avoiding that noise generated by dropping features degrades the final results.

As a result of our proposed method, we can observe in Figure 4.1 that the activation maps [108] generated by our baseline, focus both on body parts and background, whereas Top-DB-Net focus consistently on the body with stronger activation to discriminative regions.

Contrasting our Top-DB-Net with BDB Network [14], there are three differences: (i) instead of dropping random features, our method drops only features with top (the largest) activations, which stimulates the network to maintain performance using only features

[1] We use the terms *remove* and *drop* interchangeably to indicate that a tensor region has been zeroed out.

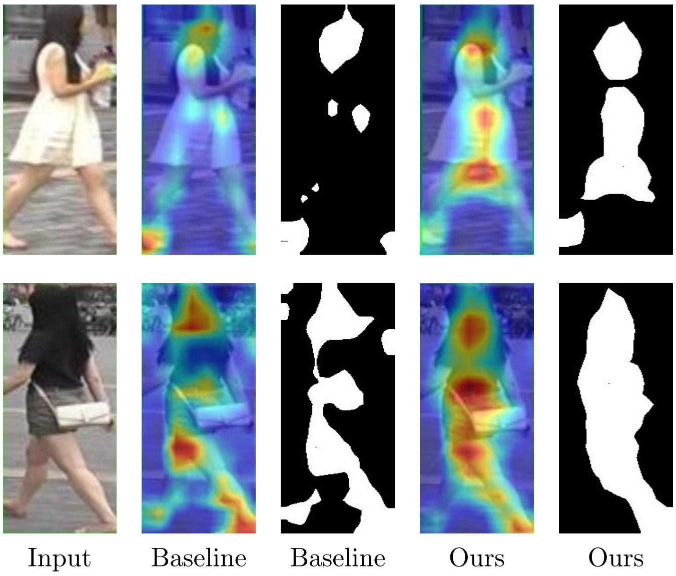


Figure 4.1: Comparison of activation maps generated by the proposed method and a baseline [14]. The first column shows the input images, the second and fourth columns present the activation maps that overlap the input images, and the third and fifth columns show a mask generated by thresholding the activation maps.

with inferior discriminative power (the lowest activations), (ii) rather than using the same drop mask for every feature map in the batch, our method creates an independent drop mask for every input based on its top activations, and (iii) dropping top activated features creates noise inside the second stream (Figure 4.2), thus we introduce a third stream that forces the features before the dropping step to be still discriminative for P-ReID, which works as a regularizer due to the multi-task principle [25].

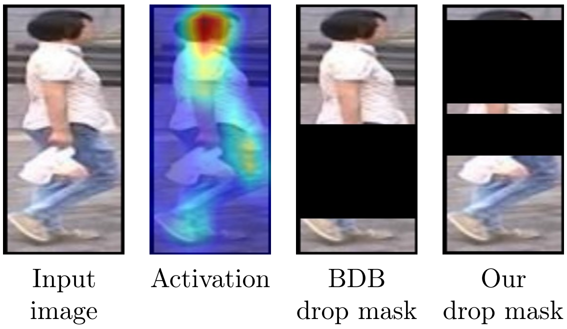


Figure 4.2: Input image, its activation map after epoch 120 and drop masks. BDB creates a random drop mask, while our method creates a mask that drops most activated regions.

We use the same definition for 'batch' as Dai et al. [14], that is, "group of images participating in a single loss calculation during training". The intuition of why our implementation is better can be explained by analyzing Figure 4.2. For an input image, we can see that the major activations are over the upper body. BDB Network [14] creates a random drop mask that, in this case, removes the lower body during training. This would encourage the network to continue focusing on the upper body. On the other hand, our method controls which regions are being dropped and encourages the network to learn from the lower body. Our results show that this helps during the learning process (Figure 4.5) and generates activation maps better spread over the foreground (Figure 4.1).

The evaluation of our proposed method is conducted through extensive experiments on three widely used datasets for P-ReID. We consider the BDB Network [14] as a baseline for our work and demonstrate that our Top-DB-Net outperforms it by up to 4.7 percentage points in the CUHK03 dataset [52]. Moreover, our results show competitive results against state-of-the-art approaches.

4.2 Baseline

We decided to use BDB Network [14] as the baseline for our proposal because of its similarity with our approach. BDB Network uses ResNet-50 [31] as backbone as in many ReID works, however, a slight variation is made by removing the last pooling layer. Thus, a larger feature map is obtained, more specifically, with a size of $2048 \times 24 \times 8$.

On top of the backbone, two streams are used. The first stream, also known as global stream, appends a global average pooling layer to obtain a 2048-dimensional feature vector. Then, a 1×1 convolution layer is used to further reduce the dimensions. The second stream, named as Batch DropBlock, *randomly* removes regions on training batches. We denote this dropping module as Batch DropBlock. Then a global maximum pooling layer is appended by creating a 2048-dimensional feature vector. A maximum pooling helps to dismiss the effect of dropped regions. Finally, a fully connected layer is used to reduce the feature vector to 1024 dimensions.

Batch DropBlock is defined to remove a region of a pre-established size based on a ratio of input images. Since BDB Network [14] reports the best results in regions with a third of height and the same width as the feature map, our Top DropBlock is defined specifically for the same scenario, this is, removing horizontal stripes.

4.3 Top-DB-Net

Our proposed network shares the same backbone as the baseline. Global, Top DropBlock and regularizer streams (Figure 4.3) are then appended. Global streams aim to extract general features directly from the backbone, following various previous approaches [14, 65, 78]. The Top DropBlock stream appends two BottleNeck layers [31] to the backbone stream and removes horizontal stripes from the most activated regions in order to push the network to maintain discriminability with less relevant data.

Given a training batch of n images, the most activated (the most informative) stripes are defined for each image independently: the backbone outputs n feature maps F of size $c \times h \times w$, where c, h and w indicates channels, height and width respectively. We transform F into an activation map A based on the definition proposed by Zagoruyko and Komodakis [108]:

$$A = \sum_{i=1}^{c} |F_i|^p \tag{4.1}$$

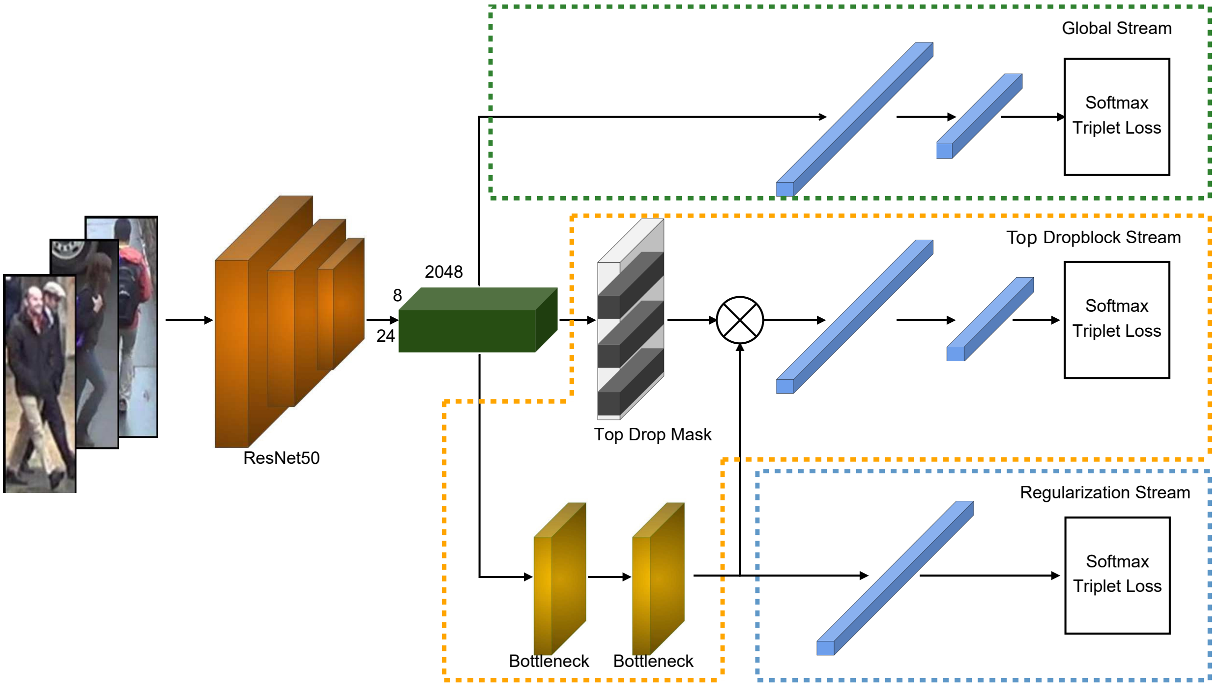


Figure 4.3: Proposed Top DropBlock Network (Top-DB-Net). It is composed of three streams that are able to focus on reliable parts of the input and encode low informative regions with high discriminative features for enhanced performance. It is trained using triplet loss and cross entropy. During the testing stage, the outputs of Global and Top DropBlock streams are concatenated.

where F_i represents every tensor slide of size $h \times w$. Assuming that $p > 1$ by definition [108], we will see that p value is not relevant to our approach.

Based on A, we define the relevance R of each stripe r_j as the average of the values on row j:

$$r_j = \frac{\sum_{k=1}^{w} A_{j,k}}{w} \tag{4.2}$$

Finally, we can zero out rows with the largest r_j values. We denote this module as Top DropBlock. For the dropping process, we create a binary mask TDM, named Top Drop Mask, of size $c \times h \times w$ for every feature map F and apply the dot product between TDM and G, where G is a tensor with the same size as F, which is the result of applying two BottleNeck layers [31] on F:

$$\text{TDM}_{i,j,k} = \begin{cases} 0, & \text{if } r_j \in \text{the largest values} \\ 1, & \text{otherwise} \end{cases} \tag{4.3}$$

such that $1 \leq i \leq c$ and $1 \leq k \leq w$.

It is worth mentioning that, from Equations 4.1 to 4.2, r_j can be expressed as:

$$r_j = \frac{\sum_{i=1}^{c} \sum_{k=1}^{w} |F^p_{i,j,k}|}{w} \tag{4.4}$$

Thus, the value of p is not relevant because $|x|^p \leq |x|^{p+1}$ for every $p > 1$ and we use r_j specifically for ranking.

Due to the $|.|$ function in the r_j definition, the most relevant stripes represent areas in F with values besides zero, both positives and negatives. We can consider those to hold more discriminative information. By removing them, we push the network to learn to distinguish between samples with less available information, thus enhancing its capabilities to encode low discriminative regions. However, if the dropped regions are too large, Top DropBlock can create noise in G due to false positives generated by removing regions that represent unique regions between different ID inputs.

To alleviate this problem, we propose a regularizer stream that will help maintain performance based on the multi-task principle [25]. This stream is only used in the training. It appends a global average pooling layer to G and is then trained for P-ReID. Thus, it encourages G to keep the information relevant to the P-ReID.

The loss function used for the three streams is the cross entropy with the label smoothing regularizer [90] and triplet loss with hard positive-negative mining [33]. During the testing process, the output of global and Top DropBlock streams are concatenated.

4.3.1 Implementation Details

All our experiments were conducted on a single Tesla v100 GPU. Due to this, we updated two items in the baseline code[2]: (i) we trained it with batch size of 64, instead of 128, and (ii) we reduced the learning rate by a factor of 0.5 × because of the "linear scaling rule" [26] to minimize the effects of training with smaller batch size.

During the training step, input images are resized to 384×128 pixels and augmented by random horizontal flip, random zooming and random input erasing [22]. As mentioned previously, our Top DropBlock stream removes horizontal stripes, thus width dropping ratio is 1. Following our baseline configuration, we use a height drop ratio of 0.3. During the testing step, no drop is applied.

Top-DropDB-Net follows the same training setup than our baseline, based on Adam Optimizer [46] and a linear warm-up [26] in the first 50 epochs with initial value of $1e-3$, then decayed to $1-e4$ and $1e-5$ after 200 and 300 epochs, respectively. The training routine takes 400 epochs. Due to the randomness of the drop masks used in our baseline and the methods used for data augmentation, we performed each experiment 5 times and reported the mean and standard deviation. This will allow for a fairer comparison with our method, baseline and ablation pipelines.

To combine cross entropy loss with label smoothing regularizer [90] and triplet loss with hard positive-negative mining [33], we used the neck method [65].

4.4 Results

In this section, we show the results for Top-DB-Net. We start with an ablation study to understand the effects of each stream and compare our Top DropBlock with Random

[2] We used author's [14] original source code available at `https://github.com/daizuozhuo/batch-DropBlock-network`

DropBlock. Then, we compare our method against the state of the art.

4.4.1 Ablation Study

We evaluate the effects of Top DropBlock and Regularization streams. Furthermore, we discuss the effects of Top DropBlock during the learning process and compare it to our baseline.

Influence of the Top DropBlock Stream

In this section, we aim to analyze the effect of our Top DropBlock stream. We train Top-DB-Net by removing the DropBlock stream and maintaining the global and regularization streams using the two Bottleneck layers. We refer to this version as “no-drop Top-DB-Net”. During testing, we concatenate the output of global and regularization branches because both streams are trained with the same loss function. Results for this comparison are shown in Table 4.1.

Table 4.1: Influence of Top-DB-Net streams and comparison with baseline.

	Market1501		**DukeMTMC-ReID**		**CUHK03 (L)**		**CUHK03 (D)**	
Method	**mAP**	**rank-1**	**mAP**	**rank-1**	**mAP**	**rank-1**	**mAP**	**rank-1**
no-drop Top-DB-Net	84.7 ± 0.1	94.4 ± 0.3	72.7 ± 0.2	86.1 ± 0.3	70.7 ± 0.4	73.8 ± 0.6	68.4 ± 0.4	71.9 ± 0.3
no-reg Top-DB-Net	83.9 ± 0.1	93.9 ± 0.2	71.1 ± 0.2	86.1 ± 0.4	71.4 ± 0.4	74.6 ± 0.8	69.4 ± 0.4	73.5 ± 1.0
Top-DB-Net	85.8 ± 0.1	94.9 ± 0.1	73.5 ± 0.2	87.5 ± 0.3	75.4 ± 0.2	79.4 ± 0.5	73.2 ± 0.1	77.3 ± 0.5
Baseline	85.2 ± 0.1	94.1 ± 0.1	73.2 ± 0.2	85.6 ± 0.3	72.2 ± 0.3	74.7 ± 0.6	70.3 ± 0.2	73.7 ± 0.4

In all datasets, we can see that removing the Top DropBlock stream decreases performance, which is also true for the standard deviation. On the Market1501 [119] and DukeMTMC-ReID [82, 121] datasets, the difference is usually less than 1 percentage points for mAP and rank-1. However, on the CUHK03 [52, 123] dataset, we can observe significant differences: performance decreases 4.7 percentage points in mAP and 5.6 percentage points in rank-1 on CUHK03(L) and decreases 4.8 percentage points in mAP and 5.4 percentage points in rank-1 on CUHK03(D). The difference in effects between datasets may be related to the fact that CUHK03 is a more challenging benchmark. This same pattern is repeated when analyzing the effect of our Regularization stream and baseline.

These results are expected because the global and regularization streams follow the same optimization logic: push the backbone to encode relevant ReID information from the input images. On the other hand, when we use our Top DropBlock stream, we further encourage the backbone to recognize relevant regions and learn to describe less informative regions with richer features.

Influence of the Regularization Stream

In this section, our goal is to show that the regularization stream, in fact, helps to deal with the noise generated by the dropping step. For this purpose, we train a version of the Top-DB-Net without this stream, named “no-reg Top-DB-Net” and compare it to

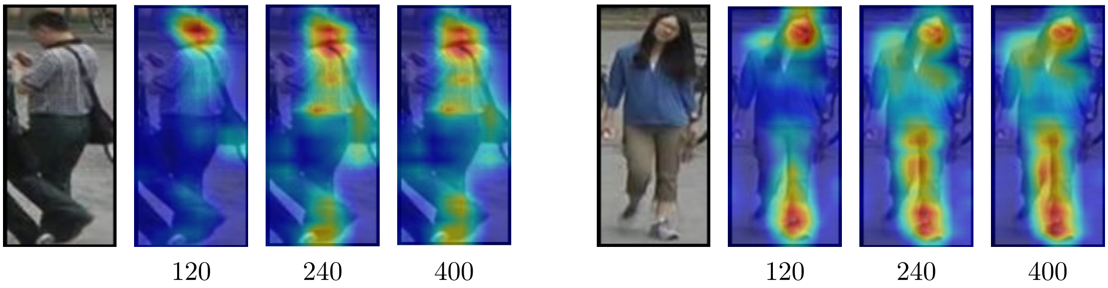


Figure 4.4: Activation maps for Top-DB-Net in two images at different epochs. The number below the images indicates the epoch.

the proposed Top-DB-Net. We can see in Table 4.1 a clear difference when using the regularization stream.

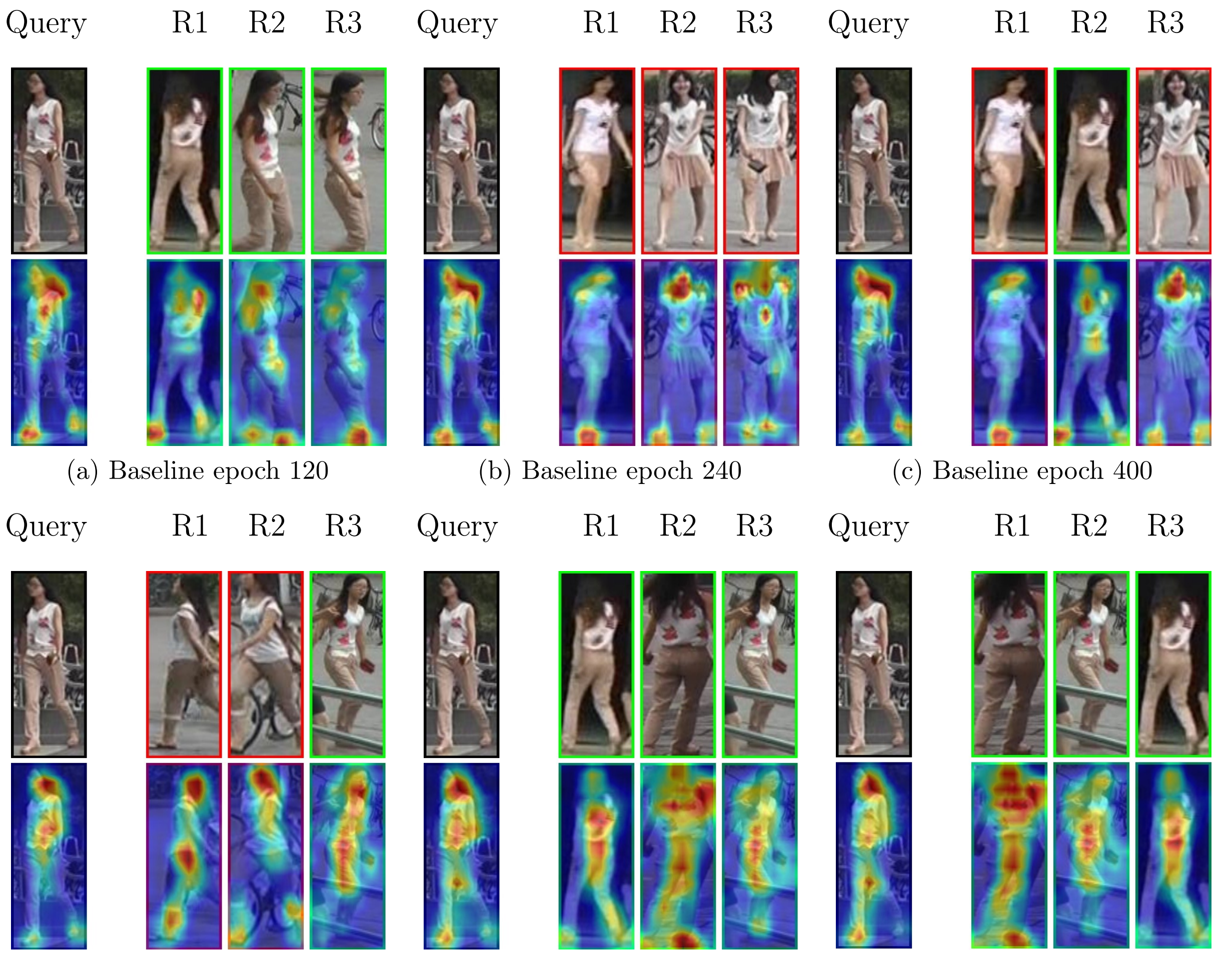
Using our regularization stream, we observe improvements of 1.9 percentage points and 1 percentage points for mAP and rank-1, respectively, on Market1501 dataset [119]. DukeMTMC-ReID [82, 121] also shows improvements of 2.4 percentage points for mAP and 1.4 percentage points for rank-1. Similar to previous ablation analysis, the most substantial changes are for CUHK03 [52, 123]: we can observe improvements of 4.8 percentage points for rank-1 and 4 percentage points for mAP on CUHK03(L), and 3.8 percentage points for rank-1 and mAP on CUHK03(D).

Random DropBlock vs Top DropBlock

Results in Table 4.1 show that our Top-DB-Net is better than our baseline in almost all metrics. The only metric with similar performance is mAP for DukeMTMC-ReID. The biggest differences are again on CUHK03 [52, 123] dataset, with up to 4.7 percentage points improvement for rank-1 and 3.2 percentage points for mAP when using our Top-DB-Net. To further understand the difference in performance, we explore activation maps and their relation with the core of our method and the baseline: DropBlocks. Figure 4.6 shows the differences between the two dropping methods.

In Figure 4.5, we compare the evolution of rank-3 at different epochs for our Top-DB-Net and baseline. We also show the evolution of the activation maps. This example shows that our Top DropBlock improves the dispersion of activation maps in the foreground and the feature extraction from images. At 120th epoch, the activations of the query are similarly spread over the upper body and feet, both in the baseline and our method. However, we can see that, in the gallery, our method is better spread across the lower body, causing Top-DB-Net to incorrectly obtain rank-1/2, confusing a person who shares pants similar to the query. At 240th epoch, we can see that the baseline activations for the query have barely changed. Moreover, because it focuses only on the upper-body and feet, it is confused with images of a person with a similar upper body, but wearing a squirt with similar color instead of pants. On the contrary, our Top-DB-Net changed its activations for the query between 120th and 240th epochs, and also focused on the lower body.

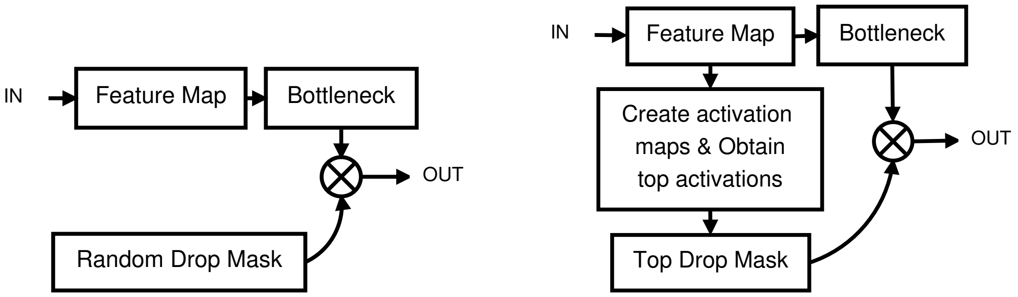
For this specific example, this is because our Top DropBlock removes the upper body regions and pushes the backbone to learn from the lower body since the 120th epoch,



(a) Baseline epoch 120 (b) Baseline epoch 240 (c) Baseline epoch 400

(d) Top-DB-Net epoch 120 (e) Top-DB-Net epoch 240 (f) Top-DB-Net epoch 400

Figure 4.5: Comparison of activation and rank-3 evolution. The top and bottom sets show images for our baseline and our proposed method, respectively. We can see that using Top DropBlock, instead of Random DropBlock, makes the activations more spread out over the person, which helps to create a better feature representation. Correct results are highlighted in green, whereas incorrect results are highlighted in red.



(a) Random DropBlock (Baseline) (b) Top DropBlock (Ours)

Figure 4.6: Differences between the Batch DropBlock and the proposed Top DropBlock.

which helps to correctly match rank-1/2/3. It is also possible to notice that, because our Top DropBlock pushes the network to describe low informative regions with rich features, at 240th epoch, our network has better features to describe lower body regions, so it fixes rank-1/2 errors of 120th epoch. Finally, at the 400th epoch, the baseline has still changed very little the distribution of its activations and still focuses only on the upper body and feet. It is able to obtain correct rank-2. On the other hand, our method still focuses on the entire body, retrieves the same correct baseline rank-2 and offers more similarity[3] to two images that show a strong viewpoint change and occlusions. This shows the improvement of the feature discriminability between 240th and 400th epochs.

Activation plots are useful for the interpretability of networks. In our case, our plots are generated following Equation 4.1. This equation is also used to define our Top DropBlock and Top Drop masks. This shows that the same tool used for interpretability can also be applied during the learning process to enhance discriminability (Table 4.1).

In addition to quantitative improvements, we observe a clear improvement in the quality of regions where backbone is concentrated. As shown in Figures 4.4 and 4.5, there is a consistent and significant activation improvement between 120th and 240th epochs, when they start to focus on broader body parts. From 240th to 400th epochs, we can see that the activations become more stable and well spread out in the foreground, but with an enhanced discriminability.

We use the activation definition by Zagoruyko and Komodakis [108] because it adjusts to our network pipeline and drop objective. However, there is a previous work for ReID [107] that uses CAM [125], an activation definition that introduces weights for each channel to enhance the scope of network activation. Previous literature and our findings suggest that methods used for interpretability may be useful to improve ReID and network activation in general.

4.4.2 Comparison with State-of-the-Art Methods

Our method focuses on ReID using information extracted only from input images. Thus, in our comparison with the state of the art, we consider methods in a similar way, for instance, Zhu et al. [131] used the camera ID and Wang et al. [97] used the image time-stamp during training. This extra information may bias the models to learn the mapping between the camera and views or the time needed for a person to move from different viewpoints, instead of extracting reliable information from images, so that they are not included in our comparison.

Table 4.2 shows a comparison between our method and state-of-the-art approaches. We compare the results separately when using re-ranking [123]. Our results are among the top-6 results for both mAP and rank-1 on Market1501. We have a similar performance for rank-1 on DukeMTMC-ReID, however, for mAP, we achieved results comparable to state-of-the-art methods, such as OSNet [128], CAMA [107] and IANet [37]. We obtained the second best rank-1 on CUHK03(L), third best mAP on both versions of CUHK03 and fourth best rank-1 on CUHK03(D). When using re-ranking, our method achieved state-of-the-art results on CUHK03(L) and CUHK03(R) in both mAP and rank-1, as well as best results for rank-1 on DukeMTMC-ReID, second best mAP on DukeMTMC-ReID and second best on Market1501 in both mAP and rank-1.

[3] Shortest Euclidean distance between features from query and gallery images.

Table 4.2: Comparison with state-of-the-art approaches. RK stands for re-ranking [123]. The sub-index indicates the ordinal position of this result (for instance, x_3 indicates that x is the third best result).

	Market1501		**DukeMTMC-ReID**		**CUHK03 (L)**		**CUHK03 (D)**	
Method	**mAP**	**rank-1**	**mAP**	**rank-1**	**mAP**	**rank-1**	**mAP**	**rank-1**
BoT [65]	85.9_5	94.5	76.4	86.4	–	–	–	–
PyrNet [66]	86.7	95.2_3	74.0	87.1	68.3	71.6	63.8	68.0
Auto-ReID [77]	85.1	94.5	–	–	73.0	77.9	69.3	73.3
MGN [96]	86.9_4	95.7_1	78.4_3	88.7_4	67.4	68.0	66.0	66.8
DenSem [115]	87.6_3	95.7_1	74.3	86.2	75.2	78.9	73.1	78.2_3
IANet [37]	83.1	94.4	73.4	87.1	–	–	–	–
CAMA [107]	84.5	94.7	72.9	85.8	–	–	–	–
MHN [5]	85.0	95.1_4	77.2	89.1_2	72.4	77.2	65.4	71.7
ABDnet [6]	88.2_2	95.6_2	78.5_2	89.0_3	–	–	–	–
SONA [104]	88.6_1	95.6_2	78.0	89.2_1	79.2_1	81.8_1	76.3_1	79.1_1
OSNet [128]	84.9	94.8	73.5	88.6_5	–	–	67.8	72.3
Pyramid [118]	88.2_2	95.7_1	79.0_1	89.0_3	76.9_2	78.9	74.8_2	78.9_2
Top-DB-Net (Ours)	85.8_6	94.9_5	73.5	87.5_6	75.4_3	79.4_2	73.2_3	77.3_4
SSP-ReID+RK [78]	90.8	93.7	83.7	86.4	77.5	74.6	75.0	72.4
BoT+RK [65]	94.2_1	95.4	89.1_1	90.3_2	–	–	–	–
PyrNet+RK [66]	94.0	96.1_1	87.7	90.3_2	78.7_2	77.1_2	82.7_2	80.8_2
Auto-ReID+RK [77]	94.2_1	95.4	–	–	–	–	–	–
Top-DB-Net+RK (Ours)	94.1_2	95.5_2	88.6_2	90.9_1	88.5_1	86.7_1	86.9_1	85.7_1

4.5 Final Considerations

In this chapter, we introduced the Top-DB-Net for person ReID. Person DeID datasets have less than two thousand different IDs. In the next chapter, we study a more challenging scenario, that is, Vehicle ReID, where there are datasets with up to forty thousand different IDs.

Chapter 5

AttributeNet: Distilling Relevant Information from Attribute Labels

In this chapter, we introduce AttributeNet, a supervised method for Vehicle Re-Identification (V-ReID) that efficiently leverages vehicle attribute labels. This method shows that distilling only task-oriented attribute information is relevant for a better feature embedding.

5.1 Introduction

Recently, there is a trend to explore additional clues for better V-ReID, such as using semantic maps [68], attributes (such as type and color) [49, 76, 92, 98, 117], viewpoints [12], and vehicle parts [12, 63, 112]. In this work, we focus on the exploration of attributes to enhance the discrimination power of feature representations. Attributes are in general invariant to viewpoint changes and robust to environment alterations.

Most of the previous attribute-based works [49, 58, 62, 76, 92, 98, 117] share a common characteristic in their design: a global feature representation is extracted from an input image using a backbone network (for instance, ResNet [31]), where this feature is followed by two types of heads, one for re-identification (ReID), and the other for attribute recognition. We refer to this design as the Vanilla-Attribute Design (VAD) and illustrate a representative VAD based Network (VAN) in Figure 5.1. One direct way to use the VAD for V-ReID is to concatenate the embedding features generated from the backbone (that is, global feature) and the attribute-based modules [62, 76].

VAD aims to drive the network to learn features that are discriminative for both V-ReID and attribute recognition, where the attributes are in general invariant to viewpoint and illumination changes. However, there is a lack of effective interaction between the attribute-based branches and V-ReID branch, where the attribute modules learn features for attribute recognition but are not explicitly designed to serve for V-ReID. Wang et al. [98] explores attributes to generate attention masks, but these masks are used only to filter the information from the global feature instead of introducing the rich attribute representation into the final feature representation.

We propose Attribute Net (ANet) to enrich the interaction between the attribute

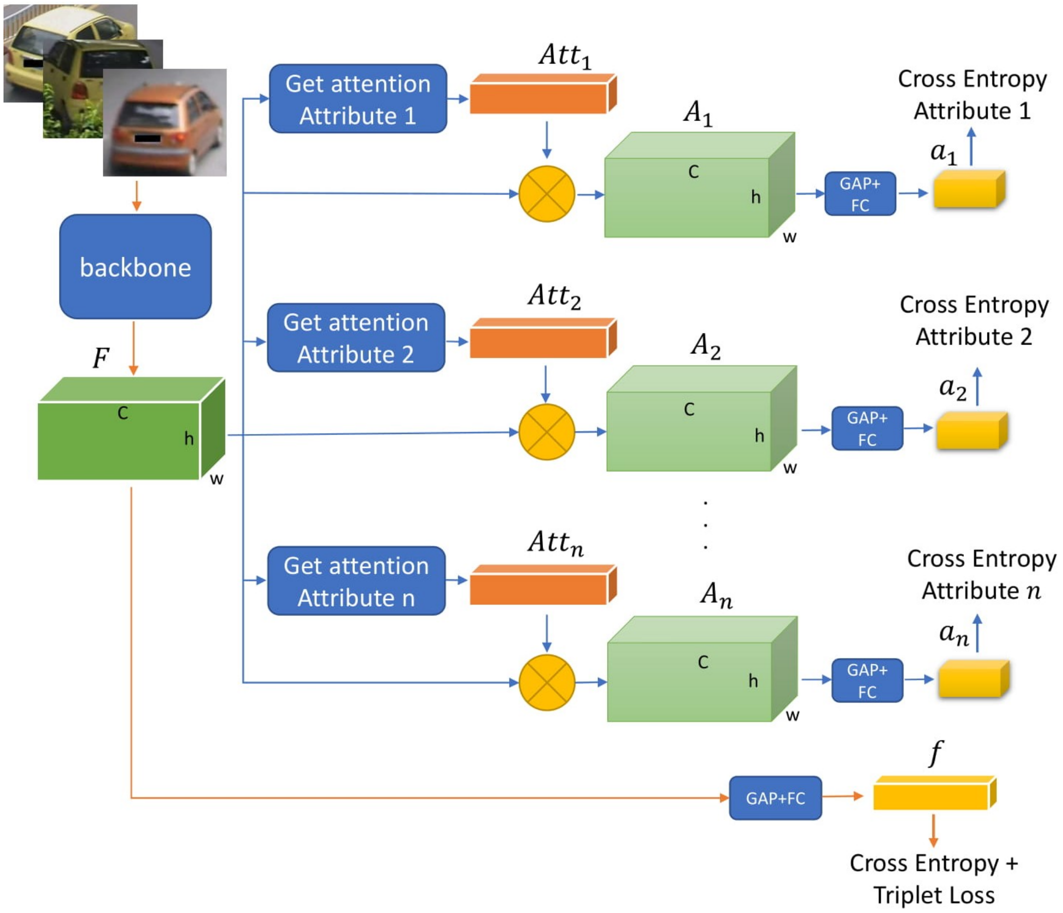

backbone
F
c
h
w
Get attention Attribute 1
Att_1
A_1
GAP+ FC
a_1
Cross Entropy Attribute 1
Get attention Attribute 2
Att_2
A_2
a_2
Cross Entropy Attribute 2
Get attention Attribute n
Att_n
A_n
a_n
Cross Entropy Attribute n
GAP+FC
f
Cross Entropy + Triplet Loss

Figure 5.1: Illustration of VAD based Network (VAN) for V-ReID. It is composed of a backbone network that learns to extract information from an input image and n branches to predict attributes based on attention modules. We use this VAN in our ANet as the first part of our framework.

features and the V-ReID feature. ANet is designed to distill attribute information and add it into the global representation (from the backbone) to generate more discriminative features. Figures 5.1 and 5.2 (with input feature maps obtained from the VAN as illustrated in Figure 5.1) present the proposed ANet. Particularly, we combine the feature maps of different attribute branches to have a unique and generic representation G of all the attributes.

We distill the helpful attribute feature from G and compensate it onto the global V-ReID feature F to have the final feature map J, where the spatial average pooled feature of J is the final ReID feature for matching. Moreover, we introduce a new supervision objective, named Amelioration Constraint (AC), which encourages the compensated V-ReID feature J to be more discriminative than the V-ReID feature F before the compensation from attribute feature.

The main contributions of this work are:

- We propose a new architecture, named ANet, for effective V-ReID, which enhances the interaction between the attribute-supervised modules and V-ReID branch. This encourages the distilled attribute features to serve for V-ReID.

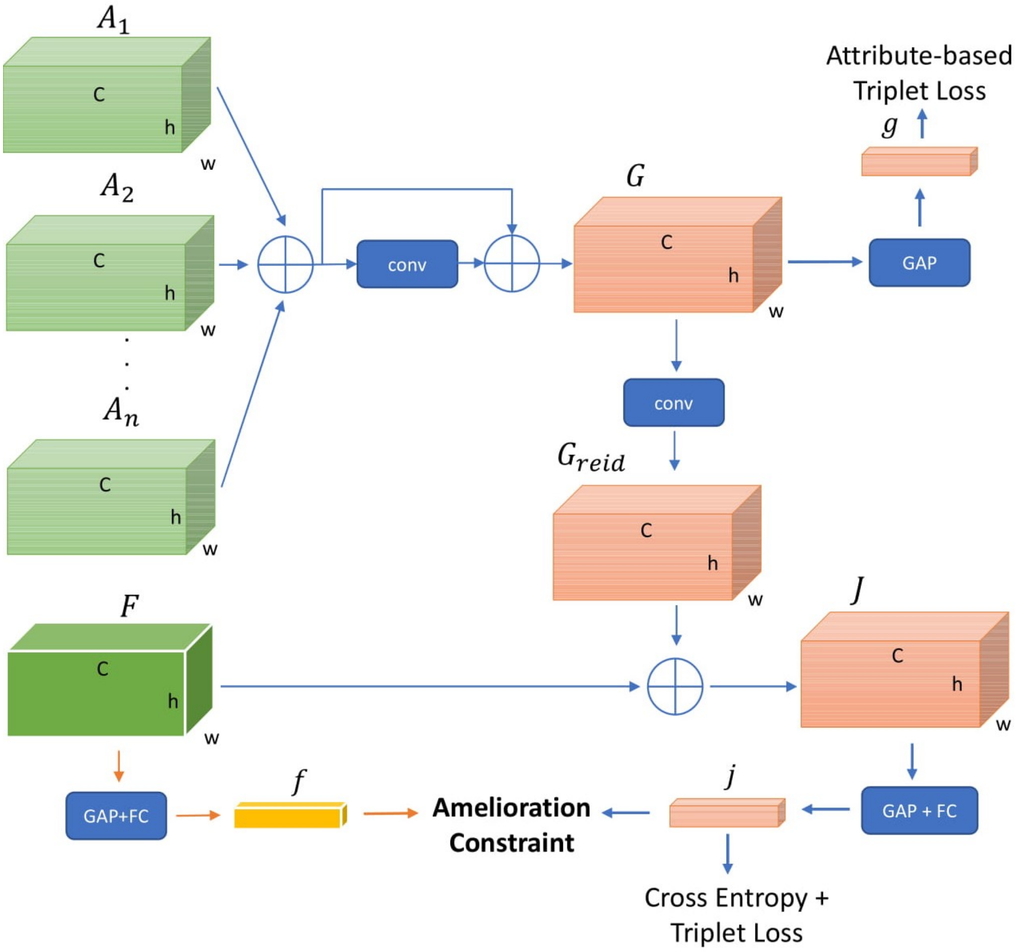


Figure 5.2: Illustration of the Joint Module. Note that the network to extract feature maps F, A_1, $\cdots$, A_n is shown in Figure 5.1 and is not shown here. We distill the helpful attribute feature from G and compensate it onto the global V-ReID feature F to have the final feature map J, where the spatial average pooled feature of J is the final ReID feature for matching. Moreover, we introduce a new supervision objective, named Amelioration Constraint (AC), which encourages the compensated V-ReID feature J to be more discriminative than the V-ReID feature F before the compensation from attribute feature.

- We introduce an Amelioration Constraint (AC), which encourages the attribute compensated feature to be more discriminative than the V-ReID feature before compensation.

Experiments on three challenging datasets demonstrate the effectiveness of our ANet, which outperforms baselines significantly and achieves state-of-the-art performance.

5.2 AttributeNet

Our proposed AttributeNet (ANet) is designed to exploit attribute information for effective V-ReID. In previous works that use attributes, there is a lack of interaction between the global V-ReID head and the attribute regression heads, which makes that the feature information is not effectively exploited for V-ReID.

To address this issue, we propose ANet (as shown in Figures 5.1 and 5.2). It consists of two parts: VAD based Network (VAN) and Joint Module (JM). VAN is based on a Backbone with two heads, where one of them is to learn global V-ReID features and the other to regress attributes. VAN outputs an initial feature representation of V-ReID and multiple Attribute features from the input image. Then, the JM distills V-ReID-helpful attribute information and compensates it into the global features. JM promotes the interaction between the attribute branches and V-ReID branch. Furthermore, we propose an Amelioration Constraint (AC), which encourages the attribute compensated feature to be more discriminative than the original V-ReID feature before the compensation.

VAD-based Network

VAD based Network (VAN), shown in Figure 5.1, aims to learn V-ReID features and regress attributes. This design is similar to previous literature work, where the attribute branches are expected to drive the learning of robust features since the attributes are in general invariant to illumination, viewpoints, etc.

Backbone. A backbone network is used to extract feature map $F(I) \in \mathbb{R}^{h \times w \times c}$ from an input image I, where h, w and c are height , width and channels of $F(I)$, respectively. We follow the previous works and use ResNet [31] as the backbone.

V-ReID Head/Branch. On top of the backbone feature $F(I)$, we append a spatial global average pooling (GAP) layer followed by a fully-connected (FC) layer to generate the V-ReID feature $f(I)$ as

$$f(I) = W_f \cdot \mathrm{GAP}\left(F(I)\right) + \mathbf{b_f}, \tag{5.1}$$

where W_f and $\mathbf{b_f}$ denote the weights and bias of the FC layer used to reduce the dimension of the pooled feature, $W_f \in \mathbb{R}^{s_f \times c}$ and $\mathbf{b_f} \in \mathbb{R}^{s_f}$, where s_f is the predefined dimension of the output. $f(I)$ is followed by Triplet Loss L^f_{tri} and Cross Entropy Loss L^f_{ID}.

Attribute Heads/Branches. On top of the backbone feature $F(I)$, we add n attribute branches for attribute classification, where n is the number of available attributes in the training dataset, one branch for each attribute. For the i-th attribute branch, we use a spatial and channel attention module to obtain attribute-related feature $A_i(I) \in \mathbb{R}^{h \times w \times c}$ as

$$A_i(I) = F(I) \cdot Att_i(F(I)), \tag{5.2}$$

where $Att_i(I) \in \mathbb{R}^c$ denotes the response of the attention module.

To make classification for the i-th attribute, we apply GAP and a FC layer to get a feature vector a_i as

$$a_i(I) = W_{a_i} \cdot \mathrm{GAP}\left(A_i(I)\right) + \mathbf{b_{a_i}}, \tag{5.3}$$

where W_{a_i} and $\mathbf{b_{a_i}}$ denote the weights and bias of the FC layer, $W_{a_i} \in \mathbb{R}^{s_a \times c}$ and $\mathbf{b_{a_i}} \in \mathbb{R}^{s_a}$, where s_a is the predefined size of the output. $a_i(I)$ is followed by a classifier with a cross entropy loss L^i_{Att} to recognize which class it belongs to for the i-th attribute.

In summary, VAN is trained by minimizing the loss L_{VAN} as

$$L_{\text{VAN}} = L^{f}_{tri} + L^{f}_{\text{ID}} + \lambda_A \sum_{i=1}^{n} L^{i}_{Att}, \tag{5.4}$$

where λ_A is a hyper-parameter for balancing the importance of V-ReID loss and attribute-related losses.

Joint Module

The Joint Module (JM) is illustrated in Figure 5.2. JM aims to distill V-ReID helpful information from the attribute features and compensate it to the V-ReID feature for the final feature matching. First, we merge the attribute feature maps from multiple branches to have a unified attribute feature map $G(I)$. Then, we distill discriminative V-ReID helpful information from $G(I)$ and compensate it onto $F(I)$ to create a Joint Feature $J(I)$. To encourage a higher discriminative capability of the Joint Feature, we introduce an Amelioration Constraint (AC), which drives the distillation of discriminative information from $G(I)$ to enhance the original V-ReID feature $F(I)$. The JM promotes the interaction between the attribute and V-ReID information to improve the V-ReID performance.

Attribute Feature $G(I)$. To facilitate the distillation of helpful attribute features, we combine all the attribute feature maps $A_i(I)$, where $i = 1, \cdots, n$, to have a unified attribute feature map $G(I)$. We achieve this by summarizing the attribute feature maps followed by a convolution layer and a residual connection as

$$G(I) = \sum_{i=1}^{n} A_i(I) + \theta_A(\sum_{i=1}^{n} A_i(I)), \tag{5.5}$$

where θ_A is implemented by a 1×1 convolutional layer followed by batch normalization (BN) and ReLU activation, that is, $\theta_A(\mathbf{x}) = \text{ReLU}(\text{W}_\text{A}\mathbf{x})$, $W_A \in \mathbb{R}^{c\times c}$. We omit BN to simplify the notation.

For the combined attribute feature map $G(I)$, we add supervision from attributes to preserve the attribute information. Given n attributes, m_i is the number of classes for the i-th attribute. There are in total $\prod_{i=1}^{n} m_i$ attribute patterns. We apply a GAP layer on $G(I)$ to get the feature vector $g(I)$. Then, the Triplet Loss L^{g}_{tri} is used as supervision to pull the features for the same attribute pattern and push the features for the different attribute patterns. We name this supervision as Attribute-based Triplet Loss.

Joint Feature $J(I)$. To distill V-ReID-helpful attribute information from $G(I)$ to enhance $F(I)$, we use two convolution layers to have distilled feature $G_{reid}(I)$

$$G_{reid}(I) = \theta_{g1}(\theta_{g2}(G(I))), \tag{5.6}$$

where θ_{g1} and θ_{g2} are implemented similarly to θ_A but we use a 3×3 convolutional layer instead of 1×1, $\theta_{g1}(\mathbf{x}) = \text{ReLU}(\text{W}_{\text{g1}}\mathbf{x})$, $\theta_{\text{g2}}(\mathbf{x}) = \text{ReLU}(\text{W}_{\text{g2}}\mathbf{x})$, $W_{g1} \in \mathbb{R}^{c\times c}$ and $W_{g2} \in \mathbb{R}^{c\times c}$.

By adding $G_{reid}(I)$ onto the V-ReID feature $F(I)$, we have the Joint Feature $J(I)$ as

$$J(I) = F(I) + G_{reid}(I). \tag{5.7}$$

$J(I)$ combines V-ReID information from $F(I)$ and the relevant V-ReID-helpful information from the attributes $G(I)$. Similar to the supervision on $F(I)$, we add Triplet Loss L^j_{tri} and Cross Entropy Loss L^j_{ID} on the spatially average pooled feature $j(I)$, where $j(I)$ is obtained as

$$j(I) = W_j \cdot \mathrm{GAP}\left(J(I)\right) + \mathbf{b_j}, \tag{5.8}$$

where W_j and $\mathbf{b_j}$ represent the weights and bias of a FC layer, $W_j \in \mathbb{R}^{s_j \times c}$ and $\mathbf{b_j} \in \mathbb{R}^{s_j}$, s_j is the predefined dimension of the output. JM is trained by minimizing L_{JM}

$$L_{\mathrm{JM}} = L^j_{tri} + L^j_{\mathrm{ID}} + \lambda_G L^g_{tri}, \tag{5.9}$$

where λ_G is a hyperparameter balancing the importance of the compensated V-ReID loss and the attribute related loss.

Finally, we can train the entire network ANet end-to-end by minimizing L

$$L = L_{\mathrm{JM}} + \lambda L_{\mathrm{VAN}}, \tag{5.10}$$

where λ is a hyperparameter to balance the importance of L_{JM} and L_{VAN}. **Amelioration Constraint.** To further boost the capabilities of the network, we define the Amelioration Constraint (AC). AC aims to explicitly encourage $j(I)$ to be more discriminative than $f(I)$. We separately apply AC for cross entropy loss and triplet loss.

AC for Cross Entropy Loss: For image I, we define it as

$$\mathrm{AC_{ID}}(I) = \mathrm{softplus}(L^j_{\mathrm{ID}}(I) - L^f_{\mathrm{ID}}(I)), \tag{5.11}$$

where $\mathrm{softplus}(\cdot) = \ln(1 + \exp(\cdot))$ is a monotonically increasing function that helps to reduce the optimization difficulty by avoiding negative values [44]. $L^f_{ID}(I)$ and $L^j_{\mathrm{ID}}(I)$ represent the identity cross entropy loss with respect to feature $f(I)$ and $j(I)$, respectively. Minimizing $\mathrm{AC_{ID}}(I)$ encourages the network to have a lower classification error for $j(I)$ than that for $f(I)$.

AC for Triplet Loss: We seek $j(I)$ to represent an enhanced feature of $f(I)$, where $j(I)$ has a higher discriminative capability than $f(I)$. Thus, we encourage the feature distance $D(\cdot, \cdot)$ between an anchor sample/image I and a positive sample I^+ to be smaller w.r.t. feature $j(\cdot)$ than feature $f(\cdot)$. Similarly, we encourage the feature distance $D(\cdot, \cdot)$ between an anchor sample/image I and a negative sample I^- to be larger w.r.t. feature $j(\cdot)$ than feature $f(\cdot)$. Then, AC for triplet loss AC_{tri} is defined as

$$\begin{aligned} AC_{tri}(I) = &\ \mathrm{softplus}(D(j(I), j(I^+)) - D(f(I), f(I^+)))+ \\ &\ \mathrm{softplus}(D(f(I), f(I^-)) - D(j(I), j(I^-))). \end{aligned} \tag{5.12}$$

We notice that training with $AC_{\mathrm{ID}}, AC_{tri}$ in an end-to-end leads to unstable learning. Thus, we follow two steps in training. In the first step, we minimize L. In the second step, we freeze the backbone (that is, all operations before f) and minimize L'. Compared with L in (Equation 5.10), the AC losses are enabled and the losses on feature f are disabled in L' as

$$L' = L + AC_{tri} + AC_{\text{ID}} - \lambda(L^f_{tri} + L^f_{\text{ID}}). \tag{5.13}$$

5.3 Implementation Details

In order to implement the backbone for a fair comparison, we follow other works available in the literature. We use a modified version of ResNet-50 [31] with Instance-Batch Normalization [73] and remove the last pooling layer to obtain the feature map $F(I)$ for an image I. Each attention module $Att_i(I)$ is based on SE [38] with the reduction ratio of 16. For the FC layers, we set $s_a = 128$ and $s_f = s_j = 512$.

We use cross entropy loss with label smoothing regularize [90] and triplet loss with hard positive-negative mining [33], following the Bag-of-Tricks [65]. For simplicity, we set $\lambda = 1$, $\lambda_A = 1$, $\lambda_G = 1$ and give the same importance to all branches in the network.

In one of the datasets, not all input images have attribute labels. For these samples, we simply do not backpropagate the losses from L^i_{Att} and L^g_{tri}. We found this works well since we use batch size of 512 (4 images per ID) and the missing labels are alleviated by the other IDs in the batch. Note that these missing labels do not affect our AC$_{\text{ID}}$ and AC$_{tri}$, so ANet can still learn from those cases.

The input images are resized to 256×256 pixels and augmented by random horizontal flipping, random zooming and random input erasing [22, 126, 127, 128]. All models are trained on 8 v100 GPUs with NVLink for 210 epochs with Amsgrad. An initial learning rate is set to 0.0006 and the learning rate is decayed by 0.1 at epochs 60, 120 and 150. The first learning step minimizes L for the first 150 epochs, then the second step optimizes L' for 60 epochs. $n = 2$ for all datasets, where we consider vehicle color (for instance, red, yellow, gray, etc.) and type (for instance, sedan, truck, etc.). During testing, the feature vectors are L2-normalized for matching.

5.4 Results

In this section, we show the results for AttributeNet. We start with an ablation study to understand the benefits of using attributes in V-ReID and how distilling task-oriented features is better than using the whole attribute information. Then, we compare our method against the state of the art.

5.4.1 Ablation Study

Our ablation study contains four subsections. In the first two subsections, we analyze the effectiveness of our ANet and its components (i.e., Joint Module and AC). In the third subsection, we aim to analyze the design of previous methods using attributes represented by VAN. Specifically, we analyze the effects of using attributes to define attention masks, instead of the FC layers, which are common in existing works. The final subsection studies the influence of hyperparameters λ_A, λ_G and λ, as well as IBN and LS.

Effectiveness of using Attributes on V-ReID

We first evaluate the effects of using attributes in V-ReID and show the comparisons in Table 5.1. *Baseline* denotes the scheme which generates feature f using only the backbone, without using attribute-related designs. VAN denotes the vanilla scheme that explores attributes as shown in Figure 5.1, using the same backbone as *Baseline*. For our VAN, we can use the V-ReID feature $f(i)$ (i.e., VAN(f)), or use the concatenation of $f(I)$ and attribute features $a_i(I), i = 1, \cdots, n$ (i.e., VAN(fa)) in inference. We can see that: (i) VAN(f), where the attributes regularize the feature learning, outperforms *Baseline* significantly on Vehicle-ID and VeRi-Wild. Specially, using attributes improves the rank-1 by 0.5 percentage points for VeRi776, 2.8 percentage points at rank-1 and 3.3 percentage points at rank-5 for Vehicle-ID, 6.6 percentage points in mAP and 1.3 percentage points at rank-1 for VeRi-Wild; (ii) using VAN(fa) has lower performance than VAN(f). This is because not all the attribute information $a_i(I)$ is equally important for V-ReID. Allocating the relative contributions of each attribute is needed to have satisfactory results. Hence how to distill task-oriented attribute information to efficiently benefit V-ReID is important, which is what our ANet aims to address.

Table 5.1: Ablation study on the effectiveness of our designs. We indicate the feature vector used for testing using the symbol in parenthesis.

	VeRi776		**Vehicle-ID**						**VeRi-Wild**					
			Small		**Medium**		**Large**		**Small**		**Medium**		**Large**	
Method	**mAP**	**R1**	**R1**	**R5**	**R1**	**R5**	**R1**	**R5**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**
Baseline	78.1	96.1	81.3	94.4	77.7	90.6	75.8	88.5	78.1	94.6	72.2	92.5	64.0	88.7
VAN (f)	78.1	96.6	84.1	96.5	80.4	93.6	78.4	91.8	83.1	94.5	78.3	93.5	70.6	90.0
VAN (fa)	77.3	96.5	81.5	95.0	78.5	92.0	76.3	89.6	81.9	94.1	76.9	93.1	69.2	89.4
ANet (j) w/o AC	79.8	96.9	85.0	96.7	80.9	94.1	79.0	91.8	84.6	96.1	79.9	94.4	72.9	91.5
ANet (j)	80.1	96.9	86.0	97.4	81.9	95.1	79.6	92.7	85.8	95.9	81.0	94.5	73.9	91.6

VAN: Attention vs Fully Connected

We use VAN as our *attribute-based* baseline, which is similar to previous works that explore vehicle attributes. However, previous works commonly used simple FC layers, instead of attention blocks for the attribute branches. Using attention facilitates the distillation of attribute features. As shown in Table 5.2, attention outperforms the use of FC layers by 1.2 percentage points in rank-1 on Vehicle-ID, as well as 1.4 percentage points and 1 percentage points in mAP on VeRi776 and VeRi-Wild datasets, respectively.

ANet: A Superior Way to Distill Attributes Information

We propose ANet to distill attribute information for more effective V-ReID. Here, we study the effectiveness of our Joint Module design and the AC losses. Table 5.1 shows the

Table 5.2: Comparison of choice for implementation of attribute branches for the attribute-based baseline VAN. *fc* represents an implementation using fully connected layers and *att* represents an implementation based on SE attention blocks. Results for Vehicle-ID and VeRi-Wild are reported using their small scale test set.

	VeRi776		**Vehicle-ID**		**VeRi-Wild**	
Method	**mAP**	**R1**	**R1**	**R5**	**mAP**	**R1**
fc	76.7	95.8	83.3	96.0	82.1	94.3
att	78.1	96.6	84.1	96.5	83.1	94.5

comparisons. We can see that: (i) our final scheme ANet (j) significantly outperforms the basic network VAN (f), by 2.0 percentage points in mAP on VeRi776, 1.9 percentage points/1.5 percentage points/1.5 percentage points in Rank-1 on Small/Medium/Large scales of Vehicle-ID, 2.7 percentage points/2.7 percentage points/3.3 percentage points in mAP on Small/Medium/Large scales of VeRi-Wild; (ii) our proposed AC losses, which encourages higher discrimination after the compensation of distilled attribute feature than that before, is very helpful to promote the distill of discriminative information from attribute feature for V-ReID purpose.

These results show that the interaction between the V-ReID and attribute features of VAN improves the network performance, thanks to the distillation of V-ReID oriented attribute features.

To better understand the effects of ANet, we visualize the attention maps of $G(I)$ and $G_{reid}(I)$ and show some in Figure 5.3. $G(I)$ encodes generic features of the attributes, where the activations are flatter and do not have a special focus on the vehicle parts. In contrast, $G_{reid}(I)$ represents a portion of the information of $G(I)$ that is helpful for V-ReID. We can observe that the activation maps focus more on the vehicle.

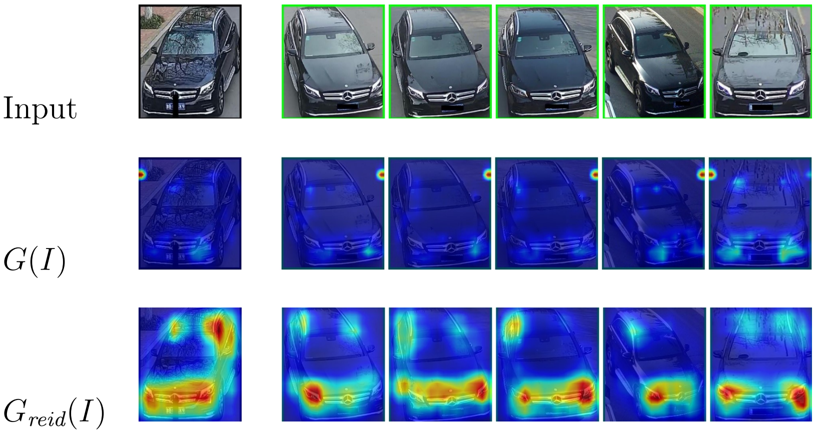


Figure 5.3: Comparison of activation maps. The first row represents the input images, second and third row their corresponding activation maps for $G(I)$ (attribute features) and $G_{reid}(I)$ (attribute features oriented to V-ReID), respectively. The first column is the query image, the second to sixth columns represent the vehicle retrieved at rank-1, rank-2, rank-3, rank-4 and rank-5.

To further analyze the effectiveness of our proposed ANet, we compare our interaction design with that using attributes such as attention, which we refer to as ANet (att). In ANet (att), attention is learned based on attributes using CBAM [102] and is used to generate attribute-guided features similar to AGNet [98]. In this case, $J(I)$ is defined through Equation 5.14 as

$$J(I) = F(I) \cdot \text{CBAM}(G(I)) + F(I). \tag{5.14}$$

We compare the performance of ANet (att) with our Anet in Table 5.3. We can see that using attention directly to increase the interaction between attribute and ReID heads is not as effective as ours. Distilling the ReID-relevant information from the attribute head defined by G_{reid} provides a superior performance. Furthermore, ANet (att) has a performance similar to the simple baseline VAN.

Table 5.3: Comparison of our interaction design with that using attributes as attention. Note that the results for Vehicle-ID and VeRi-Wild are reported using their small scale test set.

	VeRi776		**Vehicle-ID**		**VeRi-Wild**	
Method	**mAP**	**R1**	**R1**	**R5**	**mAP**	**R1**
Baseline	78.1	96.1	81.3	94.4	78,1	94.6
ANet (att)	78.2	96.1	83.9	96.2	84.9	95.5
ANet	80.1	96.9	86.0	97.4	85.8	95.9

Hyperparameter Analysis

Both hyperparameters λ_A (in Equation 5.4) and λ_G (in Equation 5.9) balance the importance of the attribute information in the total loss. We study their influence and show the results in Table 5.4. We can see that assigning the same weight for both attribute and V-ReID signals (e.g., $\lambda_A = \lambda_G = 1$) provides the best results.

Interestingly, assigning a low weight to the attribute signals (e.g., $\lambda_A = 0.01$ and $\lambda_G = 0.01$) decreases their impact and results in inferior performance. Furthermore, giving high weights to attribute signals (e.g., $\lambda_A = 100$ and $\lambda_G = 100$) is better than assigning them with rather low weights. This shows the importance of the attribute information in our pipeline for V-ReID. Finally, λ (in Equation 5.10) balances the importance of VAN and JM.

We observe that a high weight to VAN (e.g., $\lambda = 100$) significantly decreases the performance, where the contribution of our JM is small. The best weight to combine VAN and JM is $\lambda = 1$. Based on this analysis, we set $\lambda_A = \lambda_G = \lambda = 1$ and use these values in the remaining experiments.

In both our baseline scheme and final scheme, we follow the common practice and use Instance Batch Normalization (IBN) and Label Smoothing (LS). Here, we study the influence of IBN and LS on the performance of our ANet and show the results in Tables 5.5 and 5.6, respectively.

Table 5.4: Ablation study on the influence of λ_A, λ_G and λ. We evaluate using VeRi776 by keeping the non-tested hyperparameters fixed. For example, in order to analyze λ_A, we set $\lambda_G = 1, \lambda = 1$.

λ_A **Results**	**0.01**	**0.1**	**1.0**	**10.0**	**100.0**
mAP	59.1	79.4	80.1	77.5	65.9
R1	88.6	96.5	97.1	96.2	91.4
R5	94.7	98.6	98.6	98.4	96.1
λ_G **Results**	**0.01**	**0.1**	**1.0**	**10.0**	**100.0**
mAP	60.3	61.7	80.1	76.7	68.1
R1	90.0	90.9	97.1	95.8	91.6
R5	95.2	95.5	98.6	98.6	96.7
λ **Results**	**0.01**	**0.1**	**1.0**	**10.0**	**100.0**
mAP	78.3	79.7	80.1	64.4	56.4
R1	96.8	96.4	97.1	92.1	87.9
R5	98.3	98.2	98.6	96.3	94.5

Table 5.5: Influence of Instance Batch Normalization (IBN). Results for Vehicle-ID and VeRi-Wild are reported using their small scale test set.

	VeRi776		**Vehicle-ID**		**VeRi-Wild**	
Method	**mAP**	**R1**	**R1**	**R5**	**mAP**	**R1**
Baseline w/o IBN	69.5	91.0	71.5	83.1	69.2	89.2
Baseline	78.1	96.1	81.3	94.4	78,1	94.6
ANet w/o IBN	78.4	96.2	84.4	96.9	85.6	95.7
ANet	80.1	96.9	86.0	97.4	85.8	95.9

Table 5.6: Influence of Label Smoothing (LS). Results for Vehicle-ID and VeRi-Wild are reported using their small scale test set.

	VeRi776		**Vehicle-ID**		**VeRi-Wild**	
Method	**mAP**	**R1**	**R1**	**R5**	**mAP**	**R1**
Baseline w/o LS	73.2	93.2	73.8	88.4	74.3	90.1
Baseline	78.1	96.1	81.3	94.4	78,1	94.6
ANet w/o LS	79.0	96.8	85.8	97.2	85.7	95.5
ANet	80.1	96.9	86.0	97.4	85.8	95.9

For IBN, we can observe a considerable decrease of 1.7 percentage points in mAP on VeRi776 when not using IBN, whereas a decrease of 0.7 percentage points in R1. For Vehicle-ID, the difference is also significant, a decrease of 2.4 percentage points for R1 and 0.5 percentage points for R5. For VeRi-Wild, not using IBN has a smaller effect than on other datasets, that is, 0.2 percentage points in both R1 and mAP.

Without using LS, there is a decrease of 1.1 percentage points in mAP on VeRi776, 1.2 percentage points in R1 on Vehicle-ID, 0.1 percentage points in mAP on VeRi-Wild, respectively. In general, we can observe that IBN has a more significant importance than LS in the final performance.

To analyze each of the weights λ_A, λ_G and λ for our loss functions, we conduct an independent analysis per parameter (for instance, in order to analyze $lambda_A$, we evaluate 5 different values for it and fix $\lambda_G = \lambda = 1$). Results are shown in Table 5.4.

5.4.2 Comparison with State-of-the-Art Methods

We compare our method with approaches that also use attributes information [42, 49, 76, 92, 98, 117]. We also compare our method with the most recent approaches that leverage clues/techniques, such as vehicle parsing maps [68], vehicle parts [30, 112], GANs [45], Teacher-Student (TS) distillation [43, 75], camera viewpoints [12, 75], and Graph Networks (GN) [63, 86]. HPGN creates a pyramid of spatial graph networks to explore the spatial significance of the backbone tensor. PCRNet [63] studies the correlation between parsed vehicle parts through a graph network. VAnet [12] learns two metrics for similar viewpoints and different viewpoints in two feature spaces, respectively.

We also compare against FastReid [32], a strong baseline network for re-identification that performs an extensive search of hyperparameters, augmentation methods, and use some architecture design tricks to achieve excellent performance. Moreover, we implemented our design on top of it by taking it as our backbone, which we named ANet + FastReid. Note that the reported results of FastReid were obtained by our running of their released code.

Tables 5.7, 5.9 and 5.8 show the comparisons on VeRi776, Vehicle-ID, and VeRi-Wild, respectively.

VeRi776. Compared with attribute-based methods (first group in Table 5.7), our scheme ANet+FastReid outperforms the best results in this group by **5.1 percentage points** in mAP; and 1.5 percentage points for rank-1 and rank-5. By comparing with methods that do not use attributes, we can see that it performs the second best in mAP, and achieves the best for rank-1 and rank-5. VKD [75] is better than ours in mAP and is inferior to ours at rank-1 and rank-5, where VKD uses camera labels in training to be viewpoint-invariant and trains a model based on the Teacher-Student framework.

Vehicle-ID. Our method outperforms attribute-based methods (first group in Table 5.9) consistently. For rank-1, our scheme ANet+FastReid outperforms the best attribute-based method by **8.2** percentage points, **4.4** percentage points and **4.9** percentage points for small, medium and large scales, respectively. When compared with methods using other clues, ours achieves the best results on the large set and competitive performance on the other sets.

VeRi-Wild. Previous attribute based methods have not yet reported results for this latest dataset. From Table 5.8, we can see that our schemes ANet and ANet+FastReid achieve the best performance in mAP.

PVEN [68] is a method based on semantic parsing to describe each vehicle view and region. It has better results on rank-1/rank-5 but it is not as competitive as in the two

Table 5.7: Comparison of our proposed method against state-of-the-art approaches on VeRi776. The first and second best results are marked in **bold** and underline, respectively.

Method	Clues	mAP	R1	R5
PAMAL [93]	attributes	45.0	72.0	88.8
MADVR [42]	attributes	61.1	89.2	94.7
DF-CVTC [117]	attributes	61.0	91.3	95.7
PAMTRI [92]	attributes	71.8	92.8	96.9
AGNet [98]	attributes	71.5	95.6	96.5
SAN [76]	attributes	72.5	93.3	97.1
StRDAN [49]	attributes	76.1	–	–
VAnet [12]	viewpoint	66.3	89.7	95.9
PRND [30]	veh. parts	74.3	94.3	98.6
UMTS [43]	TS	75.9	95.8	–
PCRNet [63]	GN + parsing	78.6	95.4	98.4
SAVER [45]	GAN	79.6	96.4	98.6
PVEN [68]	parsing	79.5	95.6	98.4
HPGN [86]	GN	80.1	96.7	–
VKD [75]	viewpoint + TS	**82.2**	95.2	98.0
Baseline	attributes	78.1	96.1	98.3
ANet (Ours)	attributes	80.1	**97.1**	**98.6**
FastReid [32]	backbone	81.0	97.1	98.3
ANet + FastReid (Ours)	attributes	81.2	96.8	98.4

Table 5.8: Comparison of our proposed method against state-of-the-art approaches on VeRi-Wild. The first and second best results are marked in **bold** and underline, respectively.

		Small			**Medium**			**Large**		
Method	**Clues**	**mAP**	**R1**	**R5**	**mAP**	**R1**	**R5**	**mAP**	**R1**	**R5**
UMTS [43]	TS	82.8	84.5	–	66.1	79.3	–	54.2	72.8	–
HPGN [86]	GN	80.4	91.3	–	75.1	88.2	–	65.0	82.6	–
PCRNet [63]	GN + parsing	81.2	92.5	–	75.3	89.6	–	67.1	85.0	–
SAVER [45]	GAN	80.9	94.5	98.1	75.3	92.7	97.4	67.7	89.5	95.8
PVEN [68]	parsing	82.5	**96.7**	99.2	77.0	**95.4**	**98.8**	69.7	**93.4**	**97.8**
Baseline	attributes	78.1	94.6	98.5	72.2	92.5	97.3	64.0	88.7	95.6
ANet (Ours)	attributes	85.8	95.9	99.0	81.0	94.5	98.1	73.9	91.6	96.7
FastReid [32]	backbone	84.8	95.7	98.9	80.0	94.5	98.1	73.2	91.5	96.7
ANet + FastReid (Ours)	attributes	**86.9**	96.5	**99.2**	**82.5**	95.2	98.3	**75.9**	92.5	97.2

previous datasets.

We observed that none of the existing methods consistently achieve the best results on all the datasets. This may be because different datasets have different main challenges. Our proposed ANet shows a more consistent state-of-the-art performance on all the datasets, thanks to the generic capabilities of attributes on V-ReID.

Table 5.9: Comparison of our proposed method against state-of-the-art approaches on Vehicle-ID. The first and second best results are marked in **bold** and underline, respectively.

		Small		**Medium**		**Large**	
Method	**Clues**	**R1**	**R5**	**R1**	**R5**	**R1**	**R5**
PAMAL [93]	attributes	67.7	87.9	61.5	82.7	54.5	77.2
AGNet [98]	attributes	71.1	83.7	69.2	81.4	65.7	78.2
DF-CVTC [117]	attributes	75.2	88.1	72.1	84.3	70.4	82.1
SAN [76]	attributes	79.7	94.3	78.4	91.3	75.6	88.3
PRND [30]	veh. parts	78.4	92.3	75.0	88.3	74.2	86.4
SAVER [45]	GAN	79.9	95.2	77.6	91.1	75.3	88.3
UMTS [43]	TS	80.9	–	78.8	–	76.1	–
PVEN [68]	parsing	84.7	97.0	80.6	94.5	77.8	92.0
PCRNet [63]	GN + parsing	86.6	**98.1**	82.2	**96.3**	80.4	94.2
VAnet [12]	viewpoint	88.1	97.2	**83.1**	95.1	80.3	92.9
HPGN [86]	GN	**89.6**	–	79.9	–	77.3	–
Baseline	attributes	81.3	94.4	77.7	90.6	75.8	88.5
ANet (Ours)	attributes	86.0	97.4	81.9	95.1	79.6	92.7
FastReid [32]	backbone	85.5	97.4	81.8	95.3	79.9	93.8
ANet + FastReid (Ours)	attributes	87.9	97.8	82.8	96.2	**80.5**	**94.6**

5.5 Final Considerations

In the last two chapters, we introduced methods for person and vehicle ReID, respectively. In the next chapter, we study how to apply these ReID features to the Multi-Object Tracking task.

Chapter 6

AReID: Rethinking Re-Identification and Occlusions for Multi-Object Tracking

6.1 Introduction

Multi-Object Tracking (MOT) is a relevant task for the computer vision community because it can be applied in various domains, for instance, smart cities, security monitoring, action recognition, crowd behavior analysis, among others.

The most common approach used by MOT is the tracking-by-detection [55, 59, 101, 113, 114]. This type of approach has two steps: detection and matching. In the first stage, we detect the objects in each frame of the video. In the second stage, we match the identities of the detections to create the final tracks. A common design pattern in this type of work is to train an end-to-end object detector considering each frame independently.

For association, a combination of Intersection-over-Union (IoU), motion and other features are then used to match the detections. It is worth mentioning that association tends to be fully unsupervised and the relation between objects in different frames is not modeled during training, but there has been recent work that uses Transformers [11, 67, 106, 111] in order to model these relationships and interactions.

Among the tracking-by-detection approaches, there is a group of approaches that leverage ReID [55, 55, 74, 101, 101, 113, 113]. All these methods follow a common design pattern, where they add a head for ReID parallel to the detection heads during training. During matching, they use the ReID features in combination with features such as IoU and motion.

However, there is a fundamental difference between MOT and ReID training. For ReID, the model input is the bounding box of the target object that has already been segmented. In contrast, for MOT, the model input is the entire scene containing the multiple target objects and considerable background. To address this difference, the current MOT literature focuses on the centroid of the target object and projects its position onto the output tensor, then uses only this centroid during training and association (Figure 6.1).

Representing ReID features focusing only on the centroids of the target objects has

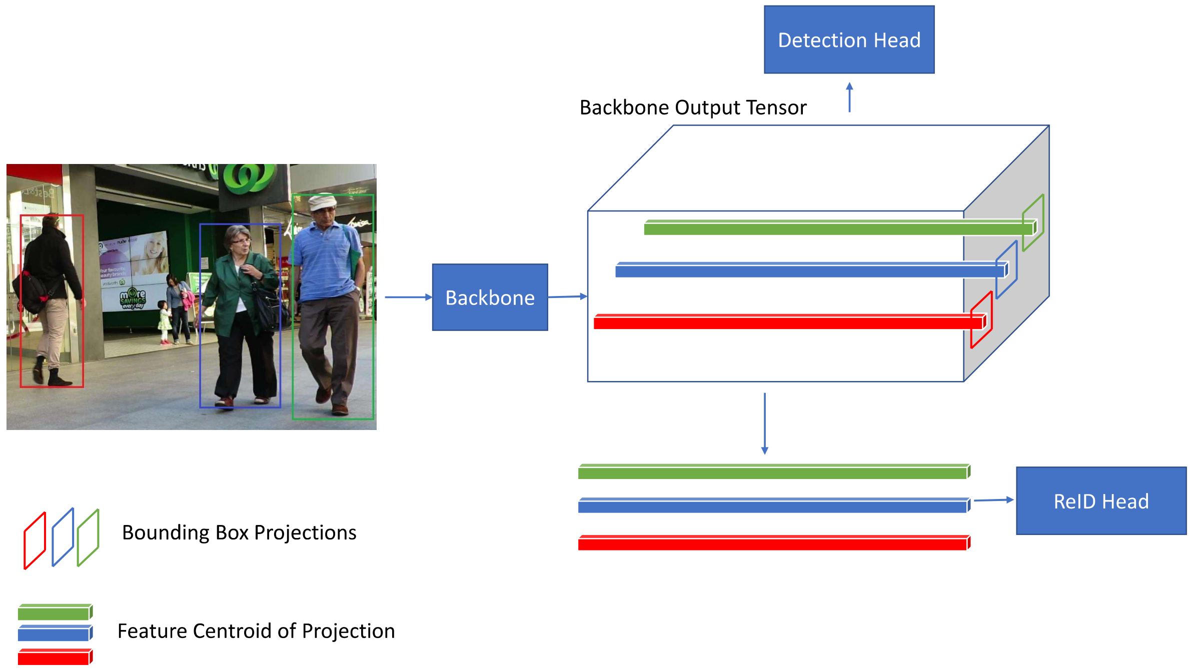


Figure 6.1: Vanilla design of ReID-based MOT methods. For the ReID head, they only use the centroid of the projections of the bounding boxes. This is different from ReID methods, where the complete backbone output tensor is used to learn ID features.

several problems that have been overlooked by the current literature, the first problem that would arise is that the centroid might not hold enough information to encode the identity of the object. In this work, we do not focus on this problem because, in MOT, the *occlusions between target objects* is a more important problem to address. In fact, our results show a regression in performance when using bounding box projection.

Occlusions between target objects is an issue for current ReID-based MOT methods because the centroid of the occluded object may be within the bounding box of the occluded object. This problem is even more common when we consider that the height and width of the output tensor of the backbone is a fraction of the input image (Figure 6.2).

In order to address this issue, we apply an adaptive use of ReID features (AReID). Specifically, before backpropagating the ReID loss of each target object, we compute the overlap between objects and give a lower weight to objects with high occlusion. This simple but effective method shows consistent improvements over various ReID-based MOT methods and datasets.

The main contributions of this approach are:

- We investigate the problem of occlusion between target objects. This has been an overlooked problem in the current literature and we conjecture that there is great potential to further boost the MOT effectiveness.

- We develop an occlusion-based weighting method that adaptively chooses the amount of ReID information that will be backpropagated.

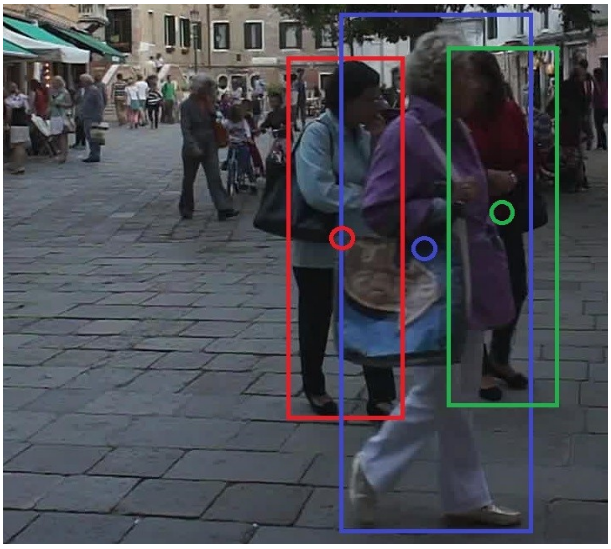


Figure 6.2: Overlap between target objects. We can see that the centroid of some objects can be occluded by other target objects. This introduces noise to the ReID head during training because it learns to represent the person ID in red and green with person features in blue.

6.2 Adaptive ReID Features

Our proposed adaptive use of ReID features (AReID) is a generic module that can be added to any ReID-based MOT approach. By using AReID, we regulate the amount of ReID information that is used during training.

The core design of previous works follows the pattern explained in Figure 6.1. However, some works may have more elaborated designs between the ReID and Detection heads. For instance, CSTrack [55] aims to disentangle ReID and Detection features, FairMOT [113] and JDE [103] have different objectives for the Detection head because of use of anchors. We will define our AReID generically and not go into the details of the Detection head. For the sake of this work, we can assume that the Detection head learns to output bounding boxes corresponding to the position of the target objects.

Backbone. A backbone network extracts the feature map $F(I) \in \mathbb{R}^{h \times w \times c}$ from an input image $I \in \mathbb{R}^{h' \times w' \times 3}$, where h, w, c are the height, width and channels of $F(I)$, respectively. In addition, h' and w' are the height and width of I. We assume that I is in the RGB space. The implementation of the network may change depending on the method. Some authors use U-Net-like networks [103, 113], while others may be based on ResNet [114].

ReID Head. Each input image I has associated a list of bounding boxes $B = \{b_1, b_2, \ldots, b_n\}$, where each $b_i = \{x_l^i, y_l^i, x_h^i, y_h^i\}, i \in \{1, ..., n\}$ is represented by the four coordinates, such that (x_l^i, y_l^i) represents the upper left corner and (x_h^i, y_h^i) represents the right lower corner of the bounding box b_i. Notice that we can calculate the centroid of b_i as

$$c_i = (c_x^i, c_y^i) = (\frac{x_l^i + x_h^i}{2}, \frac{y_l^i + y_h^i}{2}) \tag{6.1}$$

Moreover, each b_i as an associated identity id_i used to track the target object across frames.

To project c_i into $F(I)$, we have to consider the difference of size in height and width of I and $F(I)$, usually $F(I)$ is smaller than I. Let us define

$$\begin{aligned} sc_h &= \frac{h'}{h} \\ sc_w &= \frac{w'}{w} \end{aligned} \tag{6.2}$$

Then the projected centroid is expressed as:

$$c'_i = (c'^i_x, c'^i_y) = (\frac{c^i_x}{sc_w}, \frac{c^i_y}{sc_h}) \tag{6.3}$$

An important aspect here is that some works would add extra layers on top of $F(I)$ before projecting c_i. For simplicity, we do not consider this case, but the process is analogous.

Then, we calculate the centroid features $f(I)_i \in \mathbb{R}^{1\times1\times c}$ by extracting the row feature at $F(I)[c'^i_y, c'^i_x, :]$. The centroid feature is later followed by Cross Entropy Loss L_{ID}.

Adaptive ReID Weight

L_{ID} aggregates the average loss of each $f(I)_i$, which means that each centroid feature is equally considered to adjust the network weights during backpropagation. However, due to occlusions between target objects, there are cases where $f(I)_i$ is not completely reliable and introduces noise into the learning process.

Let us consider bounding boxes b_i and b_j with $i \neq j$ where b_j occludes b_i. Then, we can compute an occlusion score oc_i with $0 \leq oc_i \leq 1$ which intuitively expresses the *amount* of b_i that is occluded by b_j, a value of oc_i close to 0 means that b_i is not occluded and a value of oc_i close to 1 means b_i is completely occluded. Then, we can refine L_{ID} to adaptively consider ReID information.

$$L_{\text{ID}} = \frac{\sum\limits_{i=1}^{n} l^i_{\text{ID}}}{n} \tag{6.4}$$

where l^i_{ID} is the Cross Entropy Loss of each individual bounding box b_i. Then, we can define an adaptive L'_{ID} as:

$$L'_{\text{ID}} = \frac{\sum\limits_{i=1}^{n} g(oc_i) l^i_{\text{ID}}}{n} \tag{6.5}$$

where $g(oc_i)$ with $0 \leq g(oc_i) \leq 1$ is a function to weight the ReID features based on the amount of occlusion.

In this work, we experimented several definitions of amount of occlusion, which include IoU, a self defined asymmetric IoU, and ObjectBox [110]. We also tested various definitions of $w(oc_i)$ that include linear and hinge-like definitions. Finally, we train the network end-to-end using our newly defined L'_{ID}.

Association

The association step aims to match the identity of bounding boxes detected in two consecutive frames I and J. Let $\overset{I}{B} = \{\overset{I}{b}_1, \overset{I}{b}_2, \dots, \overset{I}{b}_n\}$ and $\overset{J}{B} = \{\overset{J}{b}_1, \overset{J}{b}_2, \dots, \overset{J}{b}_m\}$ be the set of detected bounding boxes for frames I and J, respectively. Then, we can define an affinity matrix $D \in \mathbb{R}^{n \times m}$ that indicates the cost of associating the detected objects in frames I and J, the definition of this cost is a combination of IoU, motion and ReID feature similarity. Since each bounding box in I can be associated with a unique bounding box in J, we can use D to solve the assignment problem using the Hungarian algorithm, then we will have the association with the minimum cost.

In the case of MOT, our goal is to associate the existing tracks with the detections in the new frame J. The existing tracks can be represented by the position of the last known detections (for instance, image I). Then, the process is similar to the one described previously. Notice that $n \neq m$, so we can have objects that are new in J and objects that were in I but disappear (for example, objects entering frame or occluded).

We have a mechanism that activates and deactivates tracks accordingly. An important detail is how ReID features are managed for tracks. Given the ReID feature $Feat_t$ at time t associated with a track, then its ReID feature ReID_t is defined as:

$$\text{ReID}_t = \alpha \text{ReID}_{t-1} + (1 - \alpha) Feat_t \tag{6.6}$$

where ReID_{t-1} represents the ReID feature of the track at time $t-1$. In this way the ReID features are aggregated across time creating a more robust representation as more bounding boxes are added to the track.

6.3 Implementation Details

We tested our proposed AReID in several MOT methods. For ablation studies, we used FairMOT [113] and CSTrack [55] because they are standard methods using ReID on MOT. For comparison with the state of the art, we introduced our method into ByteTrack [114], which was originally designed without ReID.

The backbone used by FairMOT [113] is a ResNet-34 combined with Deep Layer Aggregation (DLA) [129], CSTrack [55] backbone is based on Feature Pyramid Network (FPN) [57], and ByteTrack [114] backbone is based on CSPNet [99].

The Detection head used by FairMOT [113] has three parts. The first one learns to detect the centroid of the bounding box, whereas the second and third parts aim to estimate the height, width and offset of the centroid. This method is anchor free. In contrast, CSTrack is not anchor free and follows the standard objective of F-RCNN [24], that is, two coordinates of the bounding box, height and width. ByteTrack uses the PAN head [60] for detection.

The ReID heads for FairMOT [113] and CSTrack [55] composed of two convolutional layers with a 3×3 kernel size and standard ReLU and Batch Normalization layers. For ByteTrack, we added two convolutional layers to each level of the pyramid feature, which allows us to perform ReID at multiple levels. Then, we have independent Cross Entropy

Losses for each level. Notice that ByteTrack is not a new MOT method but an optimization in the association step, and the state-of-the-art results reported by the authors are based on YOLOX [21].

Input images are resized to 1088×608 for both FairMOT [113] and CSTrack [55], and 1440×800 for ByteTrack [114]. FairMOT [113] and CSTrack [55] are trained for 30 epochs with two V100 GPUs using Adam and SGD optimizers, respectively. ByteTrack [114] is trained for 60 epochs using SGD on four v100 GPUs.

We set $\alpha = 0.9$. For oc_i, we calculate the maximum amount of occlusion against any other bounding box j with $i \neq j$. We compare IoU, our non-symmetric IoU that changes the denominator of classical IoU to the area of bounding box i, and ObjectBox [110], a novel occlusion metric that considers the distance between centers and boundaries of detections. For $g(oc_i)$, we consider three functions and set $thr = 0.8$:

- Linear

$$g(oc_i) = 1 - oc_i \tag{6.7}$$

- Threshold

$$g(oc_i) = \begin{cases} 0 & \text{if } oc_i > thr \\ 1 & \text{otherwise} \end{cases} \tag{6.8}$$

- Hinge

$$g(oc_i) = \begin{cases} 0 & \text{if } oc_i > thr \\ 1 - oc_i & \text{otherwise} \end{cases} \tag{6.9}$$

All of these networks were pre-trained using MS-COCO [56]. Both MOT17 and MOT20 give only labels for the training set, and test set labels are closed. To report results against the state of the art, we use their validation server[1]. During the ablation analysis, we considered only MOT17 [69] using half of each video for training and the other half for testing. For MOT20 [15], we also followed the same protocol.

6.4 Results

In this section, we show the results of our proposed AReID. We start with an ablation study to analyze and comprehend the effect of each of the parts of our AReID. Then, we compare our method with the state of the art.

6.4.1 Ablation Study

In the first part, we analyze the effect of using the full bounding box projection versus using only the centroid for ReID feature representation. Then, we analyze the effects of different oc_i configurations. We also analyze the effects of $g(oc_i)$. Finally, after analyzing the AReID modules, we compare them to our baselines. This includes an in-depth look at the learned ReID features and qualitative analysis.

[1] `https://motchallenge.net/`

Centroid Feature for ReID

We mentioned that a problem with the current ReID design is that only the projected centroid. Hence, a natural solution to this would be to consider the projection of the bounding box followed by a pooling layer. Formally, we can define the projection $b'_i = \{x'^i_l, y'^i_l, x'^i_h, y'^i_h\}$ of bounding box b_i analogously to c'_i. Then, we define the projecting bounding box feature $f'(I)_i \in \mathbb{R}^{1\times1\times c}$ as:

$$f'(I)_i = Pool(F(I)[y'^i_l : y'^i_h, x'^i_l : x'^i_h, :]) \tag{6.10}$$

where *Pool* is the average pooling operation over the first two dimensions of $f'(I)_i$.

Table 6.1: Comparison of using centroids and full bounding box to project ReID features using FairMOT [113] and MOT17 [69] ablation setup.

Method	**IDF1**	**DR**	**DP**	**MOTA**
centroid	73.70	78.50	90.61	69.58
complete	70.95	78.20	89.78	68.37

The results of comparing centroid projected $f(I)_i$ and bounding box projected $f'(I)_i$ are shown in Table 6.1. We can see that using complete projection of bounding boxes actually leads to worse results than using only the centroid. The reason for this is the occlusion between target objects. We already explained this problem when we analyzed centroids, where it is natural to think that this problem is further amplified when using complete bounding box projection because the amount of overlap with noise is even greater, creating less reliable ReID features. Signaling this, we can see that the largest drop is in IDF1, which basically means that *complete* is worse at associating IDs.

Occlusion Definition oc_i

Here we compare different definitions for oc_i. For this comparison we fix $g(oc_i)$ to Linear, the results are shown in Table 6.2

Table 6.2: Comparison of different definitions of oc_i using FairMOT [113] and MOT17 [69] ablation setup.

Method	**IDF1**	**DR**	**DP**	**MOTA**
IoU	73.70	78.50	90.61	69.58
Asymmetric IoU	73.46	79.57	91.62	70.51
ObjectBox	73.87	79.74	91.95	70.75

We can consider IoU as the baseline for this analysis because it is the most intuitive definition for occlusion. Asymmetric IoU already gives improvements of 1.07 points for DR, 1.01 points for DP and 0.93 points in MOTA, and a small regression in IDF1. The reason for this improvement is that each oc_i is the maximum occlusion between the bounding boxes i and j, with vanilla IoU $oc_i = oc_j$ holds, but if j is the occluding object, then its score should be lower. Similarly, ObjectBox has an asymmetric definition and we can see that it pushes the IoU results even further. However, its difference from the Asymmetric IoU is less than 1 point in all metrics.

Weight Function $g(oc_i)$

In this section, we compare the three definitions of $g(oc_i)$. We fix oc_i to ObjectBox because it gives the best performance. The results are shown in Table 6.3.

Table 6.3: Comparison for different definitions of $g(oc_i)$ using MOT17 [69] ablation setup.

Method	IDF1	DR	DP	MOTA
Linear	73.87	79.74	91.95	70.75
Threshold	74.47	80.17	92.80	70.94
Hinge	73.98	80.15	92.73	71.16

We can see that the Linear definition achieves lower results more clearly in DP. However, if we consider the other metrics, the difference is in general less than 1, which is low. We consider MOT as the main metric because it considers both ID and Detection quality. Thus, Hinge is the best function for $g(oc_i)$.

AReID: Enhancing MOT with Adaptive ReID

Our proposed AReID aims to improve MOT performance by adjusting how we leverage ReID features during the training process. In this section, we compare the vanilla use of ReID information with our adaptive method. For this analysis, we decided to compare extend our validation protocol and consider FairMOT [113] and CSTrack [55] in both MOT17 [69] and MOT20 [15].

Table 6.4: Comparison of vanilla use of ReID and proposed AReID.

MOT17				
Method	**IDF1**	**DR**	**DP**	**MOTA**
FairMOT	73.70	78.50	90.61	69.58
FairMOT + AReID	73.98	80.15	92.73	71.16
CSTrack	66.24	67.91	90.72	59.57
CSTrack + AReID	66.44	69.27	90.73	60.55
MOT20				
Method	**IDF1**	**DR**	**DP**	**MOTA**
FairMOT	78.00	86.45	89.11	75.07
FairMOT + AReID	78.68	88.48	89.23	76.03
CSTrack	72.25	76.94	88.47	66.12
CSTrack + AReID	72.87	78.57	88.53	67.32

The most important conclusion from this table is that our proposed AReID gives consistent gains across datasets and methods. The largest improvements are consistently in DR (that is, up to 2.03 points), meaning that the amount of correctly detected objects by AReID is larger, which actually pushes MOTA. The largest gain in MOTA is 1.58 points for MOT17 [69], whereas the largest improvement in MOTA is 0.96 points for MOT20 [15].

In Figure 6.3, we take a closer look at the learned ReID features. When comparing detection ReID features against each other (third column) and tracking features against each other (fifth column), the perfect ReID feature would define an identity matrix, the diagonal should be red and everything else should be blue. We can see that for both cases the diagonal has the highest values. However, for our AReID, there are more regions in orange suggesting a lower quality of ReID features. This is expected because the ReID head of our method has less sample supervision as some of the samples are simply ignored.

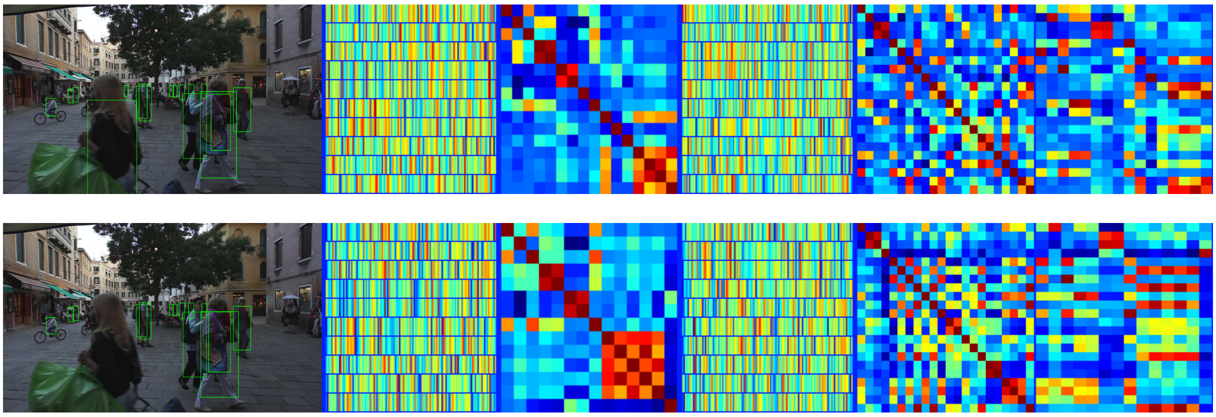


Figure 6.3: Comparison of ReID features for Vanilla (top) and proposed AReID (bottom). The first column corresponds to the detections, second column is the ReID features of the detections, the third column is the cosine distance matrix between detection ReID features, the fourth column is ReID embedding of the tracks, the fifth column is the cosine distance matrix between track ReID features, and the sixth column is the cosine distance matrix between detections and track ReID features.

If we observe the last column, the perfect ReID feature would push that each row has only one column with high value and everything else with low value (blue). We can see again that, for our AReID, there are more regions with orange. This means that our AReID overall learns less reliable ReID features than Vanilla models. However, it is no problem for MOT because, during detection, our adaptive methods help to recognize more bounding boxes (for instance, increase in DR) and during association only the maximum similarity is considered, which will ignore other near distance values.

We can also analyze in which cases Vanilla is better and in which our proposed AReID is better. Since scenes can be crowded, it is difficult to manually compare which detections are missing in each method. We then plot only the difference. In Figure 6.4, we show in red the bounding boxes correctly detected by our AReID and missed by Vanilla, in green the objects missed by our Vanilla and correctly detected by Vanilla method, and in blue the objects correctly by both method but with a large displacement.

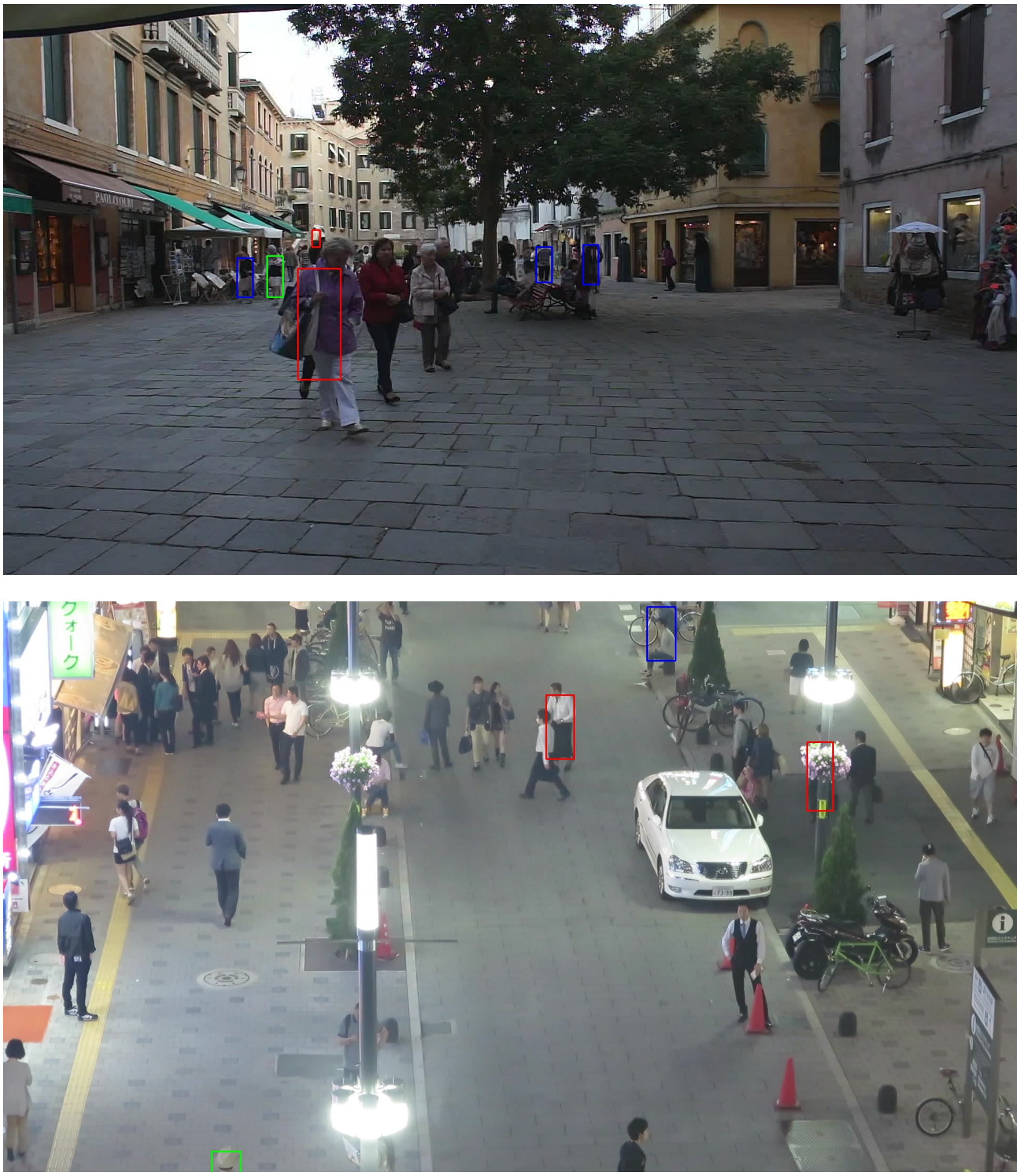


Figure 6.4: Comparison of difference in detections between Vanilla and proposed AReID. The correct detections of our AReID are shown in red, the correct detections of Vanilla in green, and detections correctly predicted but with large displacement between Vanilla and AReID are shown in blue.

We can see that our AReID is good at detecting highly occluded objects. We can also see various discrepancies in the position of the detections (for instance, in blue).

6.4.2 Comparison with State-of-the-Art Methods

In order to compare with the state of the art, we introduce our method to ByteTrack [114]. We compare our AReID with the methods using ReID features [74, 113], Transformer-based methods [67, 106, 106] and methods with the best open results on both datasets [16, 18, 36, 71].

The results for MOT17 [69] are shown in Table 6.5. The sub index indicates the ordinal position of the results. We can see that adding Vanilla ReID to ByteTrack increases by 0.4 percentage points for IDF, 0.3 percentage points for DR, 0.7 percentage points for DP and 0.4 percentage points for MOTA. However, using our AReID increases by 0.8 percentage points for IDF, 0.7 percentage points for DR, 1.8 percentage points for DP and 1 percentage points for MOTA. This shows that our method consistently improves MOT performance. Moreover, our proposed AReID achieves the best results in DR and MOTA and the top three results in IDF1 and DP.

Table 6.5: State-of-the-art comparison for MOT17 [69]. The sub index indicates the ordinal position of the results.

Method	**IDF1**	**DR**	**DP**	**MOTA**
FORTracking [71]	77.7	85.9_1	94.4	80.4
QuoVadis [16]	77.7	85.2	95.0	80.3
SelfAT [36]	79.8_1	84.7	95.0	80.0
StrongSORT [18]	79.5_2	84.7	94.5	79.6
MOTR [111]	75.0	83.2	95.3	78.6
TransCenter [106]	62.2	78.1	95.0	73.2
Quasi-Dense [74]	66.3	74.0	94.0	68.7
TraJE [23]	61.2	71.4	95.6	67.4
MPTC [88]	65.8	64.8	97.6_1	62.6
ByteTrack	77.3	85.2	95.0	80.3
ByteTrack + ReID (ours)	77.6	85.4	95.7	80.7
ByteTrack + AReID (ours)	78.1_3	85.9_1	96.8_2	81.3_1

The results for MOT20 [15] are shown in Table 6.6. Similarly to MOT17 [69], our AReID shows consistent improvements over Vanilla ByteTrack + ReID. Specifically, we show improvements of 0.7 percentage points for IDF1 and DP, 1.6 percentage points for DP and 1.2 percentage points for MOTA. Our AReID has the best results for DR and MOTA, and third best results for IDF1 and DP.

6.5 Final Considerations

In this chapter, we investigate how to apply ReID features to the Multi-Object Tracking problem. In the next chapter, we conclude the thesis with some final remarks and directions for future work.

Table 6.6: State-of-the-art comparison for MOT20 [15]. The subindex indicates the ordinal position of the results.

Method	**IDF1**	**DR**	**DP**	**MOTA**
QuoVadis [16]	75.7	83.1	94.2	77.8
SelfAT [36]	76.6_2	78.1	96.3_1	75.0
StrongSORT [18]	77.0_1	77.2	96.0_2	73.8
QDTrack [20]	73.8	79.5	94.6	74.7
FairMOT [113]	68.4	82.5	78.5	59.6
CrowdTrack [89]	68.2	75.5	94.7	70.7
TrackFormer [67]	65.7	72.9	94.9	68.6
TransCenter [106]	49.6	71.8	85.3	58.5
ByteTrack	75.2	83.1	94.2	77.8
ByteTrack + ReID (ours)	75.7	83.5	94.9	78.5
ByteTrack + AReID (ours)	75.9_3	83.8_1	95.8_3	79.1_1

Chapter 7

Conclusions

Re-Identification (ReID) is an important task that has various applications in control and surveillance of public spaces. For this reason, it has been an active research topic in the last few years. Its main challenges are inherent to the setup in which the data is collected, which includes major changes in viewpoint, illumination conditions, occlusions, among other factors. In addition to these difficulties, there are other challenges directly associated with the design of the proposed approaches, such as feature representation and invariance.

ReID can be leveraged by other tasks, for instance, Multi-Object Tracking (MOT). This task has several similar challenges to ReID. However, its main challenge is how to model simultaneous interactions and occlusions.

In this work, we presented our research in ReID considering various configurations of this problem, which includes Person ReID (P-ReID) and Vehicle ReID (V-ReID). For both cases, we focused on improving the feature representation with and without extra non-ID labels. Furthermore, we studied MOT methods that leverage ReID features to improve how ReID features are used. By splitting our work into these three studies, we were able to gain a broad understanding of ReID and its application in other tasks. In addition, our results pushed the state of the art on several fronts.

In our first approach, we explored P-ReID. We developed Top-DB-Net, a method to learn improved features from less activated regions. Top-DB-Net aims to improve overall feature representation by eliminating highly informative (for instance, highly activated) regions during training, which drives the network to learn to represent less informative regions with more reliable features. The results show improvements of up to 4.8 percentage points in mAP and 5.6 percentage points in rank-1.

In our second approach, we explored V-ReID. This setup tends to be more difficult than P-ReID because the number of objects, classes, and inter-class invariance is larger. We developed ANet, a method for distilling relevant information from attribute labels. Previous work using attribute labels has followed what we defined as the Vanilla-Attribute Design, ANet breaks this pattern and improves the interaction between the attribute-supervised branch and the ReID branch. Moreover, our proposed Amelioration Constraint pushes the performance even further. The results show that using our ANet increases mAP by up to 3.3 percentage points and rank-1 by 1.9 percentage points.

In our third approach, we explored ReID leveraged on MOT. Methods of the literature have overlooked occlusion between target class objects in their designs. We developed

AReID, a method that adaptively uses ReID features. AReID considers the maximum overlap of each bounding box and uses it to weight the ReID information. The results show an improvement in all metrics.

With the experiments presented in this thesis, we can answer the research questions introduced in Chapter 1:

Q1: Is it possible to encode rich information from areas of the image that previous methods consider less relevant for ReID?

From the experiments shown in Subsection 4.4.1, we can see that Top-DB-Net can successfully learn rich information from less activated areas. Importantly, random dropping regions instead of our proposed method leads to worse results, which makes it clear that learning to represent rich information from low activated areas is important. Moreover, we can see that Top-DB-Net tends to have a more uniform distribution of attention.

Q2: Is all the generic attribute information relevant for ReID?

Based on the experiments shown in Subsection 5.4.1, we can see that not all the attribute information is equally relevant. Specifically, by using our proposed task-relevant attribute information, we improved the mAP by 3.9 percentage points and the rank-1 by 4.5 percentage points. Furthermore, selecting task-oriented features also improves the activation maps, as shown in Figure 5.3.

Q3: How much are occlusions overlooked when applying ReID features to MOT?

The results in Subsection 6.4.1 show how occlusions between target objects are overlooked. In particular, using our AReID, we improved 2.03 percentage points in detection recall and 1.58 percentage points in MOT Accuracy. Furthermore, we showed that the effects of occlusion extend to the centroid against the argument of the full bounding box projection.

In this work, we presented novel approaches to ReID in a broad scheme. We consider that there are still several promising lines of investigation in this research topic, some of these research topics are based on follow up questions of the results shown in this thesis but others are related with some limitations of current literature. Some directions for future work are listed as follows:

- Unsupervised ReID: The cost of labeling data for ReID is high, so unsupervised learning has a great potential to reduce the total cost of deploying ReID systems. The research trend in recent years has been in this direction. However, there is still a considerable gap between supervised and unsupervised performance.

- Cross-Dataset Evaluation and Domain Generalization for ReID: In a more realistic scenario, it may be impossible to obtain images of the deployment environment before we have a functioning ReID system. Therefore, it is crucial to train models that can generalize to new data distributions. From a research point of view, this can be simulated by training and testing on different datasets or combinations of datasets and we would like the trained model to perform similarly to the model when trained and tested on the same dataset.

- Adaptive Association: Our proposed adaptive use of ReID features is primarily in the first step of tracking-by-detection. However, it makes sense to also add a similar stage during the association step to further boost results.

- End-to-End MOT: Most tracking-by-detection methods are basically supervised trained detectors with an unsupervised association step. It makes sense to join these two steps in a single end-to-end pipeline so that the training of the detector can benefit from the feedback from the association step.

- Real World ReID: An assumption of ReID definition is that for every query there is at least one element in the gallery that matches its ID. In real life, this is not always the case. Having a reliable way of finding these cases is something that has been poorly researched. Similarly, end-to-end ReID systems consist of an initial step of segmentation that has not been analyzed at the same time as ReID.

Bibliography

[1] E. Alpaydin. *Machine Learning.* MIT Press, 2021. 23

[2] P. Bergmann, T. Meinhardt, and L. Leal-Taixe. Tracking without Bells and Whistles. In *IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 20, 27

[3] G. Bonaccorso. *Machine Learning Algorithms.* Packt Publishing Ltd, 2017. 23

[4] G. Brasó and L. Leal-Taixé. Learning a Neural Solver for Multiple Object Tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020. 20, 27

[5] B. Chen, W. Deng, and J. Hu. Mixed High-order Attention Network for Person Re-identification. In *IEEE International Conference on Computer Vision*, pages 371–381, 2019. 46

[6] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but Diverse Person Re-identification. In *IEEE International Conference on Computer Vision*, pages 8351–8361, 2019. 46

[7] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person Re-identification by Multi-Channel Parts-based CNN with Improved Triplet Loss Function. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. 26, 37

[8] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom Pictorial Structures for Re-identification. In *British Machine Vision Conference*, volume 1, page 6, 2011. 26

[9] K. Chowdhary. Natural Language Processing. *Fundamentals of Artificial Intelligence*, pages 603–649, 2020. 24

[10] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, and M. D'Anastasi. Automatic Liver and Lesion Segmentation in CT using Cascaded Fully Convolutional Neural Networks and 3D Conditional Random Fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016. 23

[11] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu. TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking. *arXiv preprint arXiv:2104.00194*, 2021. 61

[12] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei. Vehicle Re-Identification with Viewpoint-aware Metric Learning. In *IEEE International Conference on Computer Vision*, pages 8282–8291, 2019. 47, 58, 59, 60

[13] E. Corvee, F. Bremond, and M. Thonnat. Person Re-identification using Haar-based and DCD-based Signature. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–8, 2010. 26

[14] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan. Batch DropBlock Network for Person Re-identification and Beyond. In *IEEE International Conference on Computer Vision*, pages 3691–3701, 2019. 9, 26, 27, 37, 38, 39, 41

[15] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. MOT20: A Benchmark for Multi-Object Tracking in Crowded Scenes. *arXiv preprint arXiv:2003.09003*, 2020. 12, 34, 35, 66, 68, 69, 71, 72

[16] P. Dendorfer, V. Yugay, A. Ošep, and L. Leal-Taixé. Quo Vadis: Is Trajectory Forecasting the Key Towards Long-Term Multi-Object Tracking? *arXiv preprint arXiv:2210.07681*, 2022. 70, 71, 72

[17] L. Deng and Y. Liu. *Deep Learning in Natural Language Processing*. Springer, 2018. 24

[18] Y. Du, Y. Song, B. Yang, and Y. Zhao. Strongsort: Make Deepsort Great Again. *arXiv preprint arXiv:2202.13514*, 2022. 70, 71, 72

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2009. 33

[20] T. Fischer, J. Pang, T. E. Huang, L. Qiu, H. Chen, T. Darrell, and F. Yu. QDTrack: Quasi-Dense Similarity Learning for Appearance-Only Multiple Object Tracking. *arXiv preprint arXiv:2210.06984*, 2022. 72

[21] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 66

[22] G. Ghiasi, T.-Y. Lin, and Q. V. Le. Dropblock: A Regularization Method for Convolutional Networks. In *Advances in Neural Information Processing Systems*, pages 10727–10737, 2018. 41, 53

[23] A. Girbau, X. Giró-i Nieto, I. Rius, and F. Marqués. Multiple Object Tracking with Mixture Density Networks for Trajectory Estimation. *arXiv preprint arXiv:2106.10950*, 2021. 71

[24] R. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 34, 65

[25] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. 24, 38, 41

[26] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv preprint arXiv:1706.02677*, 2017. 41

[27] D. Gray and H. Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *European Conference on Computer Vision*, pages 262–275. Springer, 2008. 26

[28] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu. Learning Coarse-to-Fine Structured Feature Embedding for Vehicle Re-Identification. In *AAAI Conference on Artificial Intelligence*, pages 6853–6860, 2018. 27

[29] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu. Attention Mechanisms in Computer Vision: A Survey. *Computational Visual Media*, pages 1–38, 2022. 24

[30] B. He, J. Li, Y. Zhao, and Y. Tian. Part-Regularized Near-Duplicate Vehicle Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3997–4005, 2019. 27, 58, 59, 60

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 9, 23, 24, 25, 39, 40, 47, 50, 53

[32] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei. FastReID: A Pytorch Toolbox for General Instance Re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 58, 59, 60

[33] A. Hermans, L. Beyer, and B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017. 41, 53

[34] M. Hirzer, C. Beleznai, M. Köstinger, P. M. Roth, and H. Bischof. Dense Appearance Modeling and Efficient Learning of Camera Transitions for Person Re-identification. In *IEEE International Conference on Image Processing*, pages 1617–1620, 2012. 26

[35] M. Hirzer, P. M. Roth, and H. Bischof. Person Re-identification by Efficient Impostor-based Metric Learning. In *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 203–208, 2012. 26

[36] Y. Hong, D. Li, S. Luo, X. Chen, Y. Yang, and M. Wang. An Improved End-to-End Multi-Target Tracking Method Based on Transformer Self-Attention. *arXiv preprint arXiv:2211.06001*, 2022. 70, 71, 72

[37] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Interaction-and-Aggregation Network for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9317–9326, 2019. 26, 27, 45, 46

[38] J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 53

[39] Q. Huang, W. Liu, and D. Lin. Person Search in Videos with One Portrait Through Visual and Temporal Links. In *European Conference on Computer Vision*, pages 425–441, 2018. 19

[40] Y. Huang, Q. Wu, J. Xu, and Y. Zhong. Celebrities-ReID: A Benchmark for Clothes Variation in Long-Term Person Re-Identification. In *IEEE International Joint Conference on Neural Networks*, pages 1–8, 2019. 19

[41] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, and Z. Zhang. Beyond Scalar Neuron: Adopting Vector-Neuron Capsules for Long-Term Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 19

[42] N. Jiang, Y. Xu, Z. Zhou, and W. Wu. Multi-Attribute Driven Vehicle Re-Identification with Spatial-Temporal Re-ranking. In *IEEE International Conference on Image Processing*, pages 858–862, 2018. 58, 59

[43] X. Jin, C. Lan, W. Zeng, and Z. Chen. Uncertainty-Aware Multi-Shot Knowledge Distillation for Image-Based Object Re-Identification. In *AAAI Conference on Artificial Intelligence*, 2020. 58, 59, 60

[44] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang. Style Normalization and Restitution for Generalizable Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2020. 52

[45] P. Khorramshahi, N. Peri, J.-c. Chen, and R. Chellappa. The Devil is in the Details: Self-Supervised Attention for Vehicle Re-Identification. In *IEEE European Conference on Computer Vision*, 2020. 27, 28, 29, 58, 59, 60

[46] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 41

[47] V. Kumar, A. Namboodiri, M. Paluri, and C. Jawahar. Pose-aware Person Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6223–6232, 2017. 26, 37

[48] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTchallenge 2015: Towards a Benchmark for Multi-target Tracking. *arXiv preprint arXiv:1504.01942*, 2015. 34

[49] S. Lee, E. Park, H. Yi, and S. Hun Lee. StRDAN: Synthetic-to-Real Domain Adaptation Network for Vehicle Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 608–609, 2020. 27, 28, 47, 58, 59

[50] Q. Leng, R. Hu, C. Liang, Y. Wang, and J. Chen. Person Re-identification with Content and Context Re-ranking. *Multimedia Tools and Applications*, 74(17):6989–7014, 2015. 26

[51] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017. 26, 37

[52] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep Filter Pairing Neural Network for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 20, 26, 27, 32, 39, 42, 43

[53] W. Li, X. Zhu, and S. Gong. Harmonious Attention Network for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018. 26, 27

[54] X. Li, Z. Liu, S. Cui, C. Luo, C.-F. Li, and Z. Zhuang. Predicting the effective mechanical property of heterogeneous materials by image based modeling and deep learning. *Computer Methods in Applied Mechanics and Engineering*, 347, 04 2019. doi: 10.1016/j.cma.2019.01.005. 9, 24

[55] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, and W. Hu. Rethinking the Competition Between Detection and ReID in Multi-Object Tracking. *IEEE Transactions on Image Processing*, 31:3182–3196, 2022. 20, 27, 28, 29, 61, 63, 65, 66, 68

[56] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in COntext. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 34, 66

[57] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 65

[58] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016. 27, 33, 47

[59] Q. Liu, D. Chen, Q. Chu, L. Yuan, B. Liu, L. Zhang, and N. Yu. Online Multi-Object Tracking with Unsupervised Re-identification Learning and Occlusion Estimation. *Neurocomputing*, 483:333–347, 2022. 61

[60] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path Aggregation Network for Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 65

[61] X. Liu, W. Liu, T. Mei, and H. Ma. Provid: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2017. 33

[62] X. Liu, S. Zhang, Q. Huang, and W. Gao. Ram: A Region-Aware Deep Model for Vehicle Re-Identification. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2018. 27, 47

[63] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei. Beyond the Parts: Learning Multi-view Cross-part Correlation for Vehicle Re-identification. *ACM Multimedia*, 2020. 27, 28, 47, 58, 59, 60

[64] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan. Veri-wild: A Large Dataset and a new Method for Vehicle Re-identification in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019. 34

[65] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of Tricks and a Strong Baseline for Deep Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2019. 37, 39, 41, 46, 53

[66] N. Martinel, G. Luca Foresti, and C. Micheloni. Aggregating Deep Pyramidal Representations for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–11, 2019. 46

[67] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer. Trackformer: Multi-Object Tracking with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022. 20, 27, 61, 70, 72

[68] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z.-J. Zha, X. Gao, S. Wang, and Q. Huang. Parsing-based View-aware Embedding Network for Vehicle Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2020. 27, 47, 58, 59, 60

[69] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A Benchmark for Multi-Object Tracking. *arXiv preprint arXiv:1603.00831*, 2016. 12, 34, 35, 66, 67, 68, 69, 71

[70] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. Natural Language Processing: An Introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011. 24

[71] M. H. Nasseri, M. Babaee, H. Moradi, and R. Hosseini. Fast Online and Relational Tracking. *arXiv preprint arXiv:2208.03659*, 2022. 70, 71

[72] Z. Niu, G. Zhong, and H. Yu. A Review on the Attention Mechanism of Deep Learning. *Neurocomputing*, 452:48–62, 2021. 24

[73] X. Pan, P. Luo, J. Shi, and X. Tang. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. In *European Conference on Computer Vision*, pages 464–479, 2018. 53

[74] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu. Quasi-Dense Similarity Learning for Multiple Object Tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 164–173, 2021. 20, 27, 28, 29, 61, 70, 71

[75] A. Porrello, L. Bergamini, and S. Calderara. Robust Re-Identification by Multiple Views Knowledge Distillation. *IEEE European Conference on Computer Vision*, 2020. 58, 59

[76] J. Qian, W. Jiang, H. Luo, and H. Yu. Stripe-based and Attribute-Aware Network: A Two-Branch Deep Model for Vehicle Re-Identification. *Measurement Science and Technology*, 2020. 27, 28, 47, 58, 59, 60

[77] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang. Auto-ReID: Searching for a Part-aware ConvNet for Person Re-identification. In *IEEE International Conference on Computer Vision*, pages 3750–3759, 2019. 46

[78] R. Quispe and H. Pedrini. Improved Person Re-identification Based on Saliency and Semantic Parsing with Deep Neural Network Models. *Image and Vision Computing*, 92:103809, 2019. 26, 37, 39, 46

[79] R. Quispe and H. Pedrini. Top-DB-Net: Top DropBlock for Activation Enhancement in Person Re-Identification. In *25th International Conference on Pattern Recognition*, pages 2980–2987, Milan, Italy, Jan. 2020. 22, 26

[80] R. Quispe, C. Lan, W. Zeng, and H. Pedrini. AttributeNet: Attribute Enhanced Vehicle Re-Identification. *Neurocomputing*, 465:84–92, 2021. 19, 22, 26

[81] R. Quispe, C. Lan, Z. Zhang, W. Zeng, and H. Pedrini. Rethinking Re-Identification and Occlusions for Multi-Object Tracking. *International Conference on Computer Vision (to be submitted)*, pages 1–8, 2023. 22

[82] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision Workshops*, pages 17–35, 2016. 26, 33, 42, 43

[83] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 23

[84] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A Pose-sensitive Embedding for Person Re-identification with Expanded Cross Neighborhood Re-ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018. 26

[85] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun. Crowdhuman: A Benchmark for Detecting Human in a Crowda. *arXiv preprint arXiv:1805.00123*, 2018. 34

[86] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang. Exploring Spatial Significance via Hybrid Pyramidal Graph Network for Vehicle Re-Identification. *arXiv preprint arXiv:2005.14684*, 2020. 27, 58, 59, 60

[87] N. K. Singh, M. Khare, and H. B. Jethva. A comprehensive survey on person re-identification approaches: various aspects. *Multimedia Tools and Applications*, 81 (11):15747–15791, 2022. 18

[88] D. Stadler and J. Beyerer. Multi-pedestrian Tracking with Clusters. In *17th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–10. IEEE, 2021. 71

[89] D. Stadler and J. Beyerer. On the Performance of Crowd-Specific Detectors in Multi-Pedestrian Tracking. In *17th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–12. IEEE, 2021. 72

[90] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 41, 53

[91] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple People Tracking by Lifted Multicut and Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017. 20, 27

[92] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang. Pamtri: Pose-Aware Multi-Task Learning for Vehicle Re-Identification using Highly Randomized Synthetic Data. In *IEEE International Conference on Computer Vision*, pages 211–220, 2019. 27, 28, 47, 58, 59

[93] S. Tumrani, Z. Deng, H. Lin, and J. Shao. Partial Attention and Multi-Attribute Learning for Vehicle Re-Identification. *Pattern Recognition Letters*, 138:290–297, 2020. 59, 60

[94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017. 23, 24

[95] R. Vezzani, D. Baltieri, and R. Cucchiara. People Reidentification in Surveillance and Forensics: A Survey. *ACM Computing Surveys*, 46(2):1–37, 2013. 23, 26

[96] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning Discriminative Features with Multiple Granularities for Person Re-identification. In *ACM International Conference on Multimedia*, pages 274–282, 2018. 46

[97] G. Wang, J. Lai, P. Huang, and X. Xie. Spatial-temporal Person Re-identification. In *AAAI Conference on Artificial Intelligence*, pages 8933–8940, 2019. 45

[98] H. Wang, J. Peng, D. Chen, G. Jiang, T. Zhao, and X. Fu. Attribute-guided Feature Learning Network for Vehicle Re-identification. *arXiv preprint arXiv:2001.03872*, 2020. 27, 28, 47, 56, 58, 59, 60

[99] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, and N. Zheng. End-to-End Object Detection with Fully Convolutional Network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15849–15858, 2021. 65

[100] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019. 19

[101] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang. Towards Real-time Multi-Object Tracking. In *European Conference on Computer Vision*, pages 107–122. Springer, 2020. 29, 61

[102] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision*, pages 3–19, 2018. 56

[103] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan. Track to Detect and Segment: An Online Multi-Object Tracker. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12352–12361, 2021. 29, 63

[104] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer. Second-Order Non-Local Attention Networks for Person Re-identification. In *IEEE International Conference on Computer Vision*, pages 3760–3769, 2019. 26, 27, 46

[105] H. Xiao, W. Lin, B. Sheng, K. Lu, J. Yan, J. Wang, E. Ding, Y. Zhang, and H. Xiong. Group Re-Identification: Leveraging and Integrating Multi-Grain Information. In *ACM International Conference on Multimedia*, pages 192–200, 2018. 19

[106] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus, and X. Alameda-Pineda. Transcenter: Transformers with Dense Queries for Multiple-Object Tracking. *arXiv preprint arXiv:2103.15145*, 2021. 61, 70, 71, 72

[107] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang. Towards Rich Feature Discovery with Class Activation Maps Augmentation for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1389–1398, 2019. 45, 46

[108] S. Zagoruyko and N. Komodakis. Paying more Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representations*, pages 1–13, 2017. 37, 39, 40, 45

[109] W. Zajdel, Z. Zivkovic, and B. Krose. Keeping Track of Humans: Have I seen this Person Before? In *IEEE International Conference on Robotics and Automation*, pages 2081–2086, 2005. 23

[110] M. Zand, A. Etemad, and M. Greenspan. Objectbox: From Centers to Boxes for Anchor-free Object Detection. In *European Conference on Computer Vision*, pages 390–406. Springer, 2022. 64, 66

[111] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei. MOTR: End-to-End Multiple-Object Tracking with Transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022. 20, 27, 28, 29, 61, 71

[112] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen. Part-Guided Attention Learning for Vehicle Re-identification. *arXiv preprint arXiv:1909.06023*, 2019. 27, 47, 58

[113] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. 12, 20, 27, 28, 29, 61, 63, 65, 66, 67, 68, 70, 72

[114] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. Bytetrack: Multi-Object Tracking by Associating Every Detection Box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 20, 27, 28, 29, 61, 63, 65, 66, 70

[115] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely Semantically Aligned Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 46

[116] Z. Zhang, D. Li, J. Wu, Y. Sun, and L. Zhang. MVB: A Large-Scale Dataset for Baggage Re-identification and Merged Siamese Networks. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 84–96. Springer, 2019. 19

[117] A. Zheng, X. Lin, C. Li, R. He, and J. Tang. Attributes Guided Feature Learning for Vehicle Re-identification. *arXiv preprint arXiv:1905.08997*, 2019. 27, 28, 47, 58, 59, 60

[118] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji. Pyramidal Person Re-identification via Multi-loss Dynamic Training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8514–8522, 2019. 46

[119] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable Person Re-identification: A Benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 26, 33, 42, 43

[120] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose-invariant Embedding for Deep Person Re-identification. *IEEE Transactions on Image Processing*, 28(9):4500–4509, 2019. 26, 37

[121] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *IEEE International Conference on Computer Vision*, pages 754–3762, 2017. 26, 33, 42, 43

[122] 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*, pages 2138–2147, 2019. 26

[123] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking Person Re-identification with k-reciprocal Encoding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 9, 11, 25, 26, 27, 42, 43, 45, 46

[124] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camstyle: A Novel Data Augmentation Method for Person Re-identification. *IEEE Transactions on Image Processing*, 28(3):1176–1190, 2018. 26

[125] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 45

[126] K. Zhou and T. Xiang. Torchreid: A Library for Deep Learning Person Re-identification in PyTorch. *arXiv preprint arXiv:1910.10093*, 2019. 26, 53

[127] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Learning Generalisable Omni-Scale Representations for Person Re-Identification. *arXiv preprint arXiv:1910.06827*, 2019. 53

[128] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-Scale Feature Learning for Person Re-identification. In *IEEE International Conference on Computer Vision*, pages 3702–3712, 2019. 26, 45, 46, 53

[129] X. Zhou, D. Wang, and P. Krähenbühl. Objects as Points. *arXiv preprint arXiv:1904.07850*, 2019. 23, 65

[130] Z.-H. Zhou. *Machine Learning*. Springer Nature, 2021. 23

[131] Z. Zhu, X. Jiang, F. Zheng, X. Guo, F. Huang, X. Sun, and W. Zheng. Viewpoint-Aware Loss with Angular Regularization for Person Re-Identification. In *AAAI Conference on Artificial Intelligence*, pages 13114–13121, 2020. 45