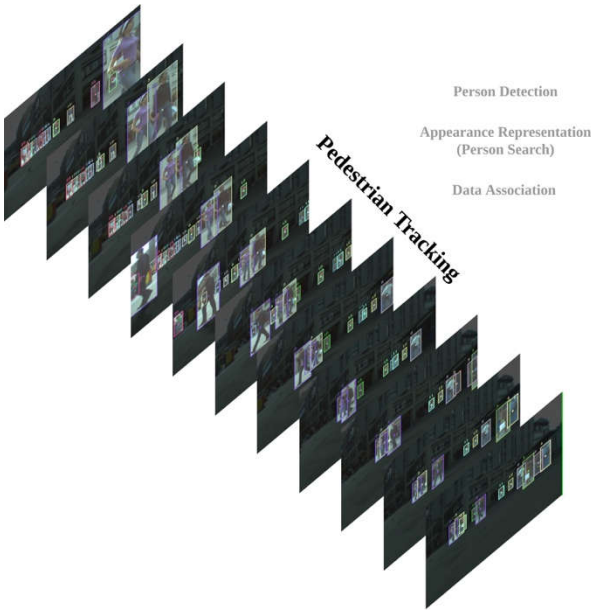
Deep Learning End-to-end Person Search and Multiple Pedestrian Tracking

Ronghua Hu

Thèse
de doctorat
de l'UTT

Ronghua HU

Deep learning End-to-end Person Search and Multiple Pedestrian Tracking



Champ disciplinaire :
Sciences pour l'Ingénieur

2021TROY0018 **Année 2021**

utt
UNIVERSITÉ DE TECHNOLOGIE
TROYES

THESE

pour l'obtention du grade de

DOCTEUR
de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

en SCIENCES POUR L'INGENIEUR

Spécialité : OPTIMISATION ET SURETE DES SYSTEMES

présentée et soutenue par

Ronghua HU

le 16 juin 2021

Deep Learning End-to-end Person Search and Multiple Pedestrian Tracking

JURY

Mme L. MERGHEM-BOULAHIA	PROFESSEURE DES UNIVERSITES	Présidente
M. Y. RUICHEK	PROFESSEUR DES UNIVERSITES	Rapporteur
M. H. TABIA	PROFESSEUR DES UNIVERSITES	Rapporteur
M. T. WANG	ASSOCIATE PROFESSOR	Examinateur
M. A. CHEROUAT	PROFESSEUR DES UNIVERSITES	Directeur de thèse
M. H. SNOUSSI	PROFESSEUR DES UNIVERSITES	Directeur de thèse

Abstract

Multi-object tracking consists of fully automated processing of image and video sequences for locating objects of interest and estimating their Spatio-temporal motion trajectories. Thanks to the rapid development of deep learning technologies, multi-object tracking and detection are used in the field of safety and security. Due to the complex appearance changes of pedestrians, non-linear motion, and mutual occlusion in crowded and mobile scenes, multiple object tracking remain extremely complex and challenging. A complete and robust tracking system consists of a detector for semantic detection, a re-identification network for pedestrians' appearance embedding representation, and an association module for trajectory maintenance and updating. In this thesis, we aim to integrate these modules using deep learning technologies for multiple object tracking by: (i) proposing of a person search network, named FT-MDnet, to extract re-identification features from multiple types of mainstream detection networks that aims at the detection, localization, and matching of pedestrians on cross-camera image galleries, (ii) proposing of a scene adaptive data association module to convert re-identification features into association features for making association decisions without the constraint of bounding boxes, and (iii) proposing of a scene adaptive detection module online feeding back the tracking result to the detection network to enhance the detection of weak and small targets.

Keywords— computer vision, machine learning, signal detection, Electronic monitoring

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Technologies Background

Ordinary RGB sensors are currently the most widely used vision sensors. Surveillance systems, robots, drones, mobile phones, and many other devices are equipped with cameras as vision sensors. Videos and images are, therefore, one of the cheapest and most accessible information materials. As a result, there is a high demand for processing and understanding these video materials, especially in some public safety and security scenes. In the past, before deep learning or, more specifically, before deep convolutional neural network enters researchers' sight, computer as the critical calculation device of computer vision plays a trivial role in understanding multiple media information, where raw images and videos always require immense human labor to participate into fetching effective semantic information. Recent years, with the emergence of deep learning technologies, image classification and object detection methods are getting remarkable developments. Cheaper and powerful GPU computing capability boosts the development of computer vision towards intelligence. Under this technical context, multiple object tracking becomes a fundamental but promising and challenging task in various computer vision applications.

Vision multiple object tracking (MOT) or multiple target tracking (MTT), plays a critical role in many applications, i.e., video surveillance, auto driving, and activity analysis. It's a challenging task which aims at fully automating image sequences and videos analysis

Figure 1.1 An illustration of multiple object tracking. Different color indicates different identities and trajectories. Best view in color and zoom in.

in order to locate, identify, and generate an inference about the motion of similar objects that belong to one or more categories without any prior knowledge about the appearance and the number of targets while maintaining their identities and yielding the individual tracks. An example of pedestrian tracking is shown in Figure. 1.1. In traditional MOT problems, the concept of 'tracking' is equivalent to data association. It is mostly applied in Radar target tracking and other point data tracking fields. Association technologies like the nearest neighbor algorithm (NNA), probability data association (PDA) [3], joint probability data association (JPDA)[4], Gaussian-mixture probability hypothesis density (GM-PHD), and multiple hypothesis tracking (MHT) [5] are applied to deal with the data association problem. Tracking filters as the Kalman filter (KF) [6] or particle filter (PF) [7] modeling system noise are widely used in many systems to fetch smooth and accurate trajectories. These traditional tracking methods have poor performances in real situations due to some domain-specific difficulties. Vision objects, especially pedestrian, because of their complex appearance changes, and non-linear motion model are hard to be detected

from complex background environments. Before using learning-based technologies in computer vision, object detection is an important preliminary step aiming at extracting some mobile foreground objects from stationary background. Nowadays, with the powerful feature representation capability of deep convolutional neural networks, better detection results have been achieved. However, when facing more difficult and complex scenes, such as crowded scenes and mobile camera situations, object detection and tracking are still challenging.

In recent years, computer vision communities continue to propose new detection methods that could provide more substantial capabilities for tackling many long-standing detection problems. Higher detection accuracy, multiple-scale object detection, instances segmentation, and occlusion object detection are coming into real practical applications. Besides object detection, person re-identification technologies focusing on learning person's appearance feature representation for cross camera person matching are widely applied in most modern multiple object trackers for providing robust inter-frame association. Data association as a cornerstone step in the tracking process is still challenging. Traditional methods like Kalman filter, Hungarian Algorithm (HA), and Kuhn-Munkres Algorithm (KMA) are dominating the data association field, but still suffer from occlusion and sudden scale variations difficulties.

MOT is a broad subject that requires incorporating object detection, appearance feature representation and data association. As a result, there have been much fewer MOT solutions that jointly consider these three application fields. Design of an end-to-end multiple object tracking systems is a long-standing challenging problem.

1.2 Research Contents and Scope

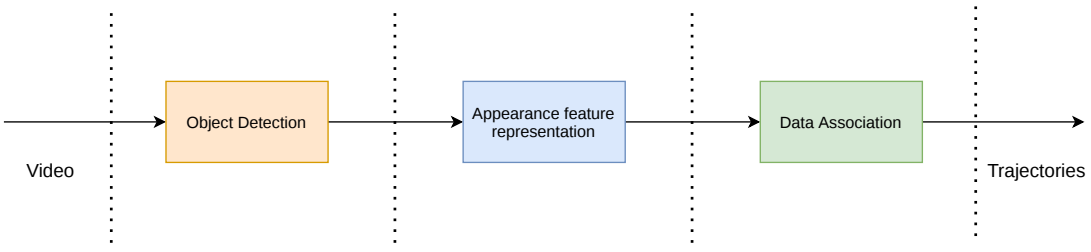


Figure 1.2 General multiple object tracking system.

To tackle the visual multiple object tracking problem, the majority of state-of-the-art tracking systems in literature are using tracking-by-detection paradigm. A general pipeline for modern tracking system is shown in 1.2. Besides object detection and data association, appearance feature representation is an essential component added into the pipeline

for robust data association in order to adapt to more complicated tracking scenes and particularly in non-overlapping cameras networks. This thesis intends to put together these three components under an uniform end-to-end framework by using deep learning techniques. The main objective is to design a single multi-branch deep neural network able to track multi-objects in overlapping/non overlapping camera networks. To this end, our work is composed of three research components, each contributing to the final whole MOT tracking system.

1.2.1 Person Search: Joint Person Detection and Re-identification across non-overlapping cameras

Pedestrian appearance feature representation is an important component in camera network pedestrian tracking. Person search is a new computer vision topic proposing an end-to-end approach for re-identification of pedestrian across non-overlapping cameras. It includes pedestrian detection and person appearance feature representation, aiming at matching a query pedestrian with gallery candidates present in given raw images. Re-identification features learned from cross-camera systems are more robust representations than intra-camera inter-frame association features. Pedestrian posture significantly varies under different camera viewpoints and thus traditional representations cannot yield discriminative features. Thus, well-trained re-identification features would focus more on the pedestrian's appearance information, such as the color and style of the pedestrian's hair, clothes, pants, and shoes. On the contrary, in intra-camera inter-frame association problem, the same pedestrian's posture change is minor, which could be treated as a main association clue. Still, it cannot provide robust association when the video frame-rate is low. In complex tracking scenes, such as mobile cameras, non-overlapping cameras or frequent loss and occlusion situations, re-identification features are critically essential to tackle these scenes. Before tackling the whole tracking system, the first part of this thesis focuses on the study of 'person search' and proposes a feasible deep-frozen transfer learning framework, named FT-MDnet to fuse multiple types of mainstream pedestrian detection networks with our proposed re-identification network, named ATLnet and MD-net, in order to extract high performance re-identification features.

1.2.2 Scene-Adaptive Data Association for intra-camera / overlapping cameras scenes

In the majority of state-of-the-art tracking algorithms, the bounding boxes generated by detectors are deemed as a main clue for inter-frame data association, as they provide

spatial constraints on the objects movements. However, in fast-moving scenes, i.e., mobile cameras or low frame-rate videos, bounding boxes between two successive frames no longer satisfy spatial constraints and the association is then prone to lose tracks. On the other hand, re-identification features (lacking of spatial constraint) are rarely used to make association decisions. They are always used as an auxiliary means to increase the accuracy of detection association. In other words, if detection association failed, the re-identification features are useless. Re-identification features, despite their discriminative capability in person search applications, are not efficient in making inter-frame association decisions in tracking applications. This thesis proposes an online-learning-scene-adaptive scheme to convert features from the re-identification feature space to an association feature space. The transformed features are trained in order to make accurate association decisions. The proposed data association is then robust and could significantly improve tracking continuity in mobile camera scenarios and long occlusion situations.

1.2.3 Feeding Back Tracking Results for Improving Detection

Despite robust re-identification features and robust data association, the whole tracking system could simply suffer from miss-detection, false alarms, detection noise, etc. This would affect the tracking performance generating trajectories fragments, false trajectories, or inaccurate target localization. In fact, in most modern tracking systems, multiple-object tracker passively receives the detector's boxes, whose performance has a decisive impact on the entire tracking system. There are only a few implementations that consider interaction between detectors and trackers. It is worth noting that a tracker connecting multiple frames detection results could confirm the targets' existence according to their spatio-temporal cues with much higher confidence than the result of a single-frame detector. By using the tracking results, it is possible to confirm whether a target is a false detection or a real object of interest. We believe that feeding back the tracking results to the detection network could improve the detection network's adaptability to the scene to a certain extent and enhance the detection performances. In this thesis, we take the detection network—CenterNet as an example to study the way of feeding back the tracking results to the detection network and conduct online training of the last few layers of the detection network to force improving the detection ability for specific targets.

1.3 Main Contributions

The main contributions of this thesis are:

1 We proposed a joint person search framework, called FT-MDnet that uses trans-

fer learning for extracting re-identification features from multiple types of mainstream detection networks. The proposed re-identification network is an end-to-end network aimed at simultaneously detect pedestrian and extract discriminative features for re-identification across non-overlapping cameras. Detection and re-identification are two different tasks, features for one task could be irrelevant for the other task. The challenge is to design an unified framework for both tasks. The proposed framework could easily include state-of-the-art detection networks, like YOLOv3, YOLOv4, Mask RCNN, and CenterNet, without the need of considering their implementation details; this helps to reduce the design difficulties. Our experimental results proved that pre-trained detection networks could preserve sufficient discriminative information for re-identification, and the fact that the feature representation of detection networks are highly compatible with the re-identification task. The re-identification network training is not from scratch but starting with these high compatible features. Thus, a shallow-designed re-identification network, like our proposed ATLnet and MDnet would work well. Moreover, our proposed framework achieved state-of-the-art performance on CUHK-SYSU and PRW, two 'person search' datasets. The top-1 and mAP search accuracy outperform all other existing person search solutions.

2 In the previous section, re-identification is aimed at finding pedestrian across different cameras without taking into account spatial constraints (the lost target could appear at any point in a new camera of the network). In this section, we consider the complementary task of tracking people inside the same camera visual scene. We argue that re-identification features play a vital role in the robust association (between successive frames) of multiple object tracking in some complex scenes. The re-identification features should be independent of the detection of bounding boxes to make association decisions in order to solve the data association problem in fast-moving or low frame-rate scenes. We propose an online-learning-scene-adaptive transformation to map the features from re-identification feature space to association feature space. The significance of association features gets remarkable improvement and has shown the advantages of association features in making association decisions. By using this method, our proposed tracker could effectively tackle the continuity tracking problem.

3 We propose a method that allows to feedback tracking results to the detection network and to online train the detection network to enhance its detection capability on the tracked objects. This method could effectively enhance the ability of detecting weak targets, thereby enhancing the continuity of tracking.

4 We comprehensively take into account the object detection, appearance feature representation, and data association problem in an end-to-end framework to design the global multiple object tracking solution. Our proposed FT-MDnet and

Online-Learning-Scene-Adaptive could be easily transplanted to other state-of-the-art detection networks and trackers to improve their data association capabilities. This flexibility to be plugged to other detector architectures is another advantage of our proposed solution.

1.4 Thesis outline

The contents of this thesis are organized as follows:

- In chapter 2, we review the previously published MOT-related works, including object detection, person re-identification, person search, multiple object tracking, and the metrics used to evaluate multiple object tracking performance.
- In chapter 3, we present our first contribution: a joint person search framework, named FT-MDnet for cross-camera pedestrian re-identification feature representation.
- In chapter 4, we details our second contribution aimed at enhancing the intra-camera inter-frame association of targets. We discuss the application of re-identification features in the data association problem and propose an online-learning-scene-adaptive transformation to convert the re-identification features into association features.
- In chapter 5, we propose a solution to feedback tracking results into the detection network in order to enhance its performances. Finally, we describe our global end-to-end tracking system in a non-overlapping camera network.
- In chapter 6, we conclude the thesis and state the perspective research work.

CHAPTER 2

Literature Review: State Of The Art

Multiple object tracking is a long-standing fundamental and challenging task in computer vision with various application scenes in public safety and security, robotics, auto driving, activity analysis, etc. In recent years, the application of deep convolutional neural networks brings a remarkable development to a number of computer vision research topics. The most MOT-related state-of-the-art technologies, like object detection, re-identification, and tracking, are updating each short season. This chapter gives a review of these MOT-related works that were published in the past few years.

2.1 A Brief Introduction of Object Detection

Figure 2.1 An example of object detection

Object detection is a fast-developing computer vision task. It is a starting point for designing multiple object tracking algorithms and almost determines the tracking per-

formance. Given an image, its task is to locate the category-specified objects and give the category information. The currently most representative one-step detector are YOLO [8, 9, 10] serials detectors. YOLO balance the inference speed and detection accuracy and truly met real-time detection requirements. They are worth-considering detection networks when concerning real-world applications. RCNN [11, 12, 13, 14] series detectors are two-step detectors, they having high detection accuracy are widely used in many vision-based research topics such as person re-identification, person search, MOT, etc.

YOLO and RCNN are anchor-based methods that exhaustively enumerate pixel-wise candidate bounding boxes and refine by performing bounding box regression. Anchor-based methods need post-processing processes, like non-maximum-suppression (NMS), to remove overlapped or duplicated bounding boxes. Recently, a more concise anchor-free concept is proposed and starting to show its advantages in dealing with the object detection problem. The most typical anchor-free detectors are CornerNet [15] and CenterNet [16].

Mainstream CNN-based object detection networks consist of at least two functional modules, terms as backbone and head. Backbone networks, like ResNet [17], Darknet [10], VGG [18], etc. are designed to extract general-purpose features from pre-training on large-scale object classification dataset like ImageNet [19]. Head is designed to fulfill specific tasks like object classification, bounding box regression, or object segmentation.

2.1.1 Two-step object detectors

Girshick et al.[11] propose the first region-based CNN detector—R-CNN, which shows that CNN-based detectors could achieve dramatic performance improvement on PASCAL VOC [20] dataset when comparing with other traditional HOG-like feature-based detectors. With the success of R-CNN, CNN-based object detection methods are verified as effective and efficient. In the latter years, RCNN serials are getting powerful with the proposal of Fast RCNN, Faster RCNN, and Mask RCNN.

— **R-CNN**[11] consists of 5 modules. **Module I** adopts a non-learning based selective search strategy[21] to give an image a coarse scan to select out around 2000 candidate region proposals. **Modele II** is a CNN structure that consists of 5 convolutional layers and 2 fully connected layers (one for classification, one for bounding box regression) that learn to extract 4096-dimensional feature vectors from cropped and resized (227×227) region proposals; all parameters in CNN are shared across all object categories. **Module III** is a set of pre-trained class-specific linear SVMs for object classification. **Module IV** is a bounding box regressor for

refining the region proposals to get precise bounding box prediction. **Module V** is a greedy non-maximum suppression function that is used to filter out duplicated region proposals.

— **Fast R-CNN**[13] is a faster version of R-CNN. Considering the CNN structure of R-CNN performing on each region proposals lead to a large amount of time and memory costs, Fast R-CNN propose a more efficient training method that takes the advantage of feature sharing during training. Besides, Fast RCNN uses a streamlined training process with a one-fine-tuning state that jointly optimizes a softmax classifier and bounding-box regressors. Fast R-CNN firstly performs the CNN once on the entire image for feature extraction. A region of interest (ROI) pooling layer is used to get the fixed-size feature regions and fed into the following classification and regression fully connected layers. Instead of using SVMs for classification, Fast RCNN jointly train the network by using multi-task loss for classification and regression. The proposed Fast RCNN saves 8× the training time of R-CNN, and the testing speed was 213× faster than R-CNN while achieving higher detection accuracy (66.0% to 66.9% mAP on PASCAL VOC 2007 dataset).

— **Faster R-CNN**[12] is an even faster version of Fast R-CNN. Fast R-CNN uses a selective search strategy to provide coarse region selection, which costs almost the same running time as the CNN part. Faster R-CNN proposes a fully convolutional network, called region proposal network (RPN), to efficiently predict region proposals with a wide range of scales and aspect ratios anchor boxes. Faster R-CNN is the first end-to-end CNN-based detection architecture. It achieved 69.9% mAP on PASCAL VOC 2007 while Faster R-CNN's total running time is nearly 10× lower than Fast RCNN. As a consequence, Faster R-CNN becomes the most popular detection network in various computer vision tasks.

— **Mask R-CNN** utilize an additional mask branch for the intra-class instance segmentation task. It is an extending work to Faster R-CNN. Mask R-CNN proposes to use a feature pyramid network (FPN)[22] to collect multiple-stage feature maps from the backbone network ResNet. The FPN contains a top-down pathway to fuse multi-scale feature maps; the generation of higher resolution feature maps are used to predict small-scale objects. Besides FPN, Mask RCNN proposes to use an aligned ROI pooling (RoIAlign) layer bilinearly interpolate and sample feature regions to avoid quantization of RoI boundaries or bins for higher accuracy of localization and bounding box regression.

2.1.2 One-step object detection

YOLO *(you only look once)* serials detectors proposed after Faster RCNN are mainly consider to improve the inference efficiency of the entire detection network. It frames object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. YOLO is the first detector that meets the requirement of real-time. Currently, YOLO family detectors include YOLO [8], YOLOv2 [9], YOLOv3 [10], YOLOv4 [23].

- **YOLO**: The backbone of YOLO containing 24 convolutional layers extracts features from a size-fixed (448×448) input image. The $64\times$ downsampled feature map has shape $7 \times 7 \times 1024$. The head of YOLO consists of two fully connected layers and outputs a $7 \times 7 \times 30$ tensor. The output tensor contains the detection information; each pixel of it includes 2 bounding boxes (x, y, w, h) indicating the location and size of the target, 2 confidence score indicating the confidence of the object fall into the specific grid, and 20 categories parameters indicating the score of the object category belonging. Last, non-maximum suppression is applied to remove duplicated detection. YOLO achieved 63.4% mAP with 45 fps on PASCAL VOC dataset.
- **YOLOv2** adopts a large number of novel concepts to improve YOLO's speed and precision. A new backbone network named Darknet-19 is proposed where batch normalization is the first time introduced in YOLO serials detectors. YOLOv2 pre-trained on ImageNet with high-resolution input images (448×448) to adapt the high-resolution requirement in the object detection task. In YOLO, bounding boxes are directly generated by fully connected layers. YOLOv2 adopts Faster R-CNN's prediction mechanism that replaces fully connected layers with convolutional layers and introduces the anchor box concept to predict and refine bounding boxes. YOLOv2 also proposed to use dimension clusters to define the size and aspect ratio of anchor boxes. Other concepts like fine-grained features were used to improve small size objects prediction, and multi-scale training used to adapt different size input images. Consequently, YOLOv2 achieved 78.6% mAP, $40 fps$ as compared to YOLO's 63.4% mAP, $45 fps$ on PASCAL VOC dataset.
- **YOLOv3** propose a backbone network named Darknet-53. The FPN-like structure is used in YOLOv3 to predict bounding boxes on different scale feature maps; this makes YOLOv3 detect small-size objects better. YOLOv3 uses multi-label classification (using binary cross-entropy instead of sigmoid for the class predictions) to adopt more complex datasets containing many overlapping labels.
- **YOLOv4** introduced a number of new concepts and training tricks, including weighted-residual-connections, cross-state-partial-connections [24] self-adversarial-training, mish-activation [25], and mosaic data augmentation, etc. to improve the

detection results to the state-of-the-art.

2.1.3 New concepts for object detection

The design of detectors is tending to be complicated for achieving better accuracy. Recently, some new concepts and new architecture were proposed that could significantly reduce the burden of designing detection networks. Strictly speaking, their methods are still one-step detection methods but more intuitive and straightforward. They showed the future development tendency of detection algorithms.

— **Anchor-Free**

CornerNet [15] facility the detection work as finding the top-left corner and bottom-right corner pairs. It produces two gaussian kernel heatmaps, two embedding feature maps, and two offset maps. The heatmaps are used to locate the top-left corners and bottom-right corners using a so-called corner pooling layer to search the peak of heatmaps. The offset maps are used to adjust the position of corner points slightly. The embedding feature maps giving embedding vectors for the corner points are used to measure the similarity and match top-left bottom-right pairs in between. Without introducing anchor boxes, the design of the loss function becomes straightforward.

CenterNet use a similar idea as CornerNet that predicts a bounding box size-related gaussian kernel heatmap and using maximum pooling layer instead of corner pooling layer to locate the center of objects. It directly regresses the object's height and weight without introducing anchor boxes; consequently, CenterNet has achieved higher detection accuracy and less inference time than YOLOv3. Moreover, directly detect the object center point make CenterNet well adapted for applying in the tracking problem.

— **Detection Transformer**

DETR [26] adopting an encoder-decoder architecture based on transformers. It simplifies the detection pipeline by dropping multiple hand-designed components that encode prior knowledge, like anchor boxes or non-maximal suppression. DETR bridges the gap between natural language processing (NLP) and computer vision by proving that a transformer could replace convolutional neural networks while keeping comparable performance.

2.2 Person Re-Identification and Person Search

Appearance re-identification features could provide non-temporal constraints association cues for tracking. Different from other tracking algorithms that mainly focus on learning inter-frame appearance representation. We argue that a robust tracker should adapt to different frame-rate videos; the appearance feature should work across multiple frames to model the sharp posture variation. Person re-identification and person search are tightly MOT-related vision topics for the cross-camera appearance representation problem. They are widely studied as a specific person retrieval problem overs non-overlapping cameras. Due to the urgent demand for public safety and security, re-identification is imperative in intelligent video surveillance system designs. There are many publications in recent years.

2.2.1 Person Re-Identification

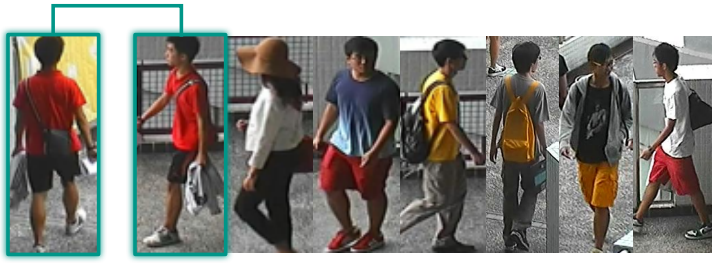


Figure 2.2 An example of person re-identification [1].

Person re-identification aims to match manually cropped pedestrian images between queries and candidates. It is a challenging task due to the appearance changes in different camera viewpoints [27, 28], varying image resolutions [29], illumination variations [30], unconstrained person posture [31] etc.. With the application of learning-based technologies, computer vision communities proposed and published many papers, standard datasets, and benchmarks for re-identification to promote the development of it. We focus on reviewing the most recent and solid learning-based works.

Multi-Part re-identification feature representation

Global feature learning directly learns and extracts feature representations on the whole person image without the part constraints. Due to the originally application of deep neural networks in image classification problem, global feature learning is a primary choice

in the early age to integrating advanced deep learning techniques [32, 33]. It is discriminative when the person detection can accurately locate the human body. When person images suffer from large background clutter or heavy occlusions, part-level feature learning usually achieves better performance by mining discriminative body regions [34, 35, 36]. The solutions of AlignedReID [37], AlignedReID++ [38], are proposed to solve the image unaligned problem. Their solutions is to split an image into 8 strips and separately learn part-level feature representation for dynamically matching local information across images. In [39] a mask-guided contrastive attention model is proposed to eliminate the impact of background clutter, trained with a region-level triplet loss. Multiple granularity network (MGN) [40] propose to use a multi-branch deep CNN architecture for global and local feature representations. The local branch consists of different scale segmentations called multiple granularity to represent different body regions of human. Due to the advantage in handing misalignment or occlusions, most of state-of-the-art methods developed recently adopt the features aggregation paradigm that combining the part-level and global-level features.

Small-Scale re-identification feature representation

In facing the increasingly challenging re-identification scenes, large-scale or global-level appearance representations, i.e., color of clothes, shoes, hair etc. are insufficient to provide discriminative capabilities. Instead, small-scale/detailed appearance feature, i.e., logos, detailed textures, are critically important because their uniqueness [41, 42, 43]. OSnet [44] propose the concept of omni-scale feature learning to mining detailed features via a network that constructed with multiple receptive scale bottlenecks.

2.2.2 Person Search

Figure 2.3 An example of person search [1].

The topic of person search is proposed for the end-to-end application of re-identification technologies. It integrating object detection and person re-identification aims to both

locate and match pedestrians on a gallery of raw images. The introducing of pedestrian detection gives domain-specific difficult to person search, making person search a even more chanllenging problem than person re-identification. By the way, person search is a more MOT-related topic since it accomplished the object detection task and appearance feature representation problem. Considering the network architecture, there are two types of person search solutions.

Independent Person Search

Independent person search (IPS) including three independent components: **i**. a pedestrian detector detecting pedestrians. **ii**. a module cropping and resizing pedestrian images. **iii**. a Re-ID network extracting discriminative features. IPS solutions, including [45, 46, 47], taking the multi-scale problem as the key to improving Re-ID accuracy, have achieved better accuracy than joint person search solutions. MGTS [48] integrating pedestrian detection, and pedestrian segmentation make an effort to tackle the background and foreground (pedestrian) cues separately. Due to its easy integration of pedestrian detection and person re-identification, IPS solutions always achieve better search accuracy than joint person search solutions.

Joint Person Search

Joint person search aim to jointly detect pedestrian and extract discriminative features from an end-to-end network. In 2014, Xu et al. [49] proposed a combination framework consists of sliding window pedestrian detection and handcrafted feature matching. Later, a number of learning based solutions were proposed [1, 50, 51, 52, 53, 54, 55]. In 2017, Xiao et al. proposed the first CNN based framework named OIM[1] that jointly implements pedestrian detection and appearance feature extraction in an end-to-end manner and achieved a top-1 accuracy of 78.7% on CUHK-SYSU, which is a person search dataset. The success of OIM inspired the latter joint person search (JPS) solutions. Liu et al. [52] studied the loss function of OIM and proposed a multi-loss fusion strategy to speed up the network convergence and compact the intra-class distance, and their approach achieved 79.9% top-1 accuracy. LCGPS [53] introduced a relative attention module to employ the scene context cues to decrease the search confusion, and it gave a top-1 accuracy of 86.5%. QEEPS [54] using a query-guided strategy to improve the feature representation between query and gallery images has achieved the highest top-1 accuracy of 89.1%. It is worth noting that these JPS solutions are still not as accurate as IPS solutions, e.g., in [45], the top-1 accuracy is 89.9%, and in [47], the top-1 accuracy is 91.7%.

2.3 Multiple Object Tracking

MOTchallenge[1] provides a benchmark for a fair comparison of the performance of trackers. A tracker using detection provided by MOTchallenge is called public detector tracker, otherwise, called private detector tracker. According to the processing mode, there are two research communities in multiple object tracking. Online tracking considers the application scene of some real-time surveillance scene that the video frames sequentially input to the tracker. The tracker is expected to update the state of targets based on current or previous video frames. Offline tracker is a post-processing tracking technology that could use all available video information (past, present, future) for optimization. Offline tracker comparing with online tracker always has better tracking continuity and trajectory localization accuracy.

There are four types of detectors as follows:

1 public detector + online tracking
2 public detector + offline tracking
3 private detector + offline tracking
4 private detector + online tracking

Usually, multiple object tracking refers to online tracking; that is, the video is available to framewise fed into the tracker for updating the tracks' states. Online data association is the core part of multiple object tracking, and it is always designed as a data assignment problem between existed tracks and next arrived detection. Data association solving on the spatial-scale and in the feature space are two main implementations.

2.3.1 Spatial-Scale Association

SORT [56] taking the detection bounding boxes as the only association clue for tackling the data association problem. SORT approximate the inter-frame displacements of each object with a linear constant velocity model which is independent of other objects and camera motion. The state of each target is modeled as a 7-dimensional vector including the horizontal and vertical pixel location of the center of the target, the scale and the aspect ratio of the target's bounding box, and their corresponding velocity except the aspect ratio. The assignment cost matrix is computed as the IoU distance between detection and predicted bounding boxes. And the solver is Hungarian algorithm.

MAT [57] argue that associate per-frame detection is difficult to ensure long-range tracking when camera motion, fast motion, and occlusion challenges occur. Re-identification

1. https://motchallenge.net

features due to lack of temporal-spatial constraints are unreliable, time-consuming, and still cannot address the false negatives for occluded and blurred objects. MAT focusing more on various motion patterns of different objects proposes an enhanced MOT paradigm, namely motion aware tracker. It improves MOT from three aspects: **i**. it blends the nonrigid pedestrian motion and rigid camera motion seamlessly to balance their compatible issues. **ii**. it consists of a general dynamic reconnection context module to ensure robustness and smoother filled tracking fragments for long-range motion-based reconnection. **iii**. it applies the temporal-spatial constraints to filter useless track-detection association connections with lower time cost by 3D integral image encoding.

DHN [58] propose a differentiable proxy of the two metrics— multiple object accuracy (MOTA) and multiple object tracking precision (MOTP) for end-to-end training of deep multiple object tracker. A key ingredient of DHN is the proposing of a Deep Hungarian Net module that approximates the Hungarian algorithm and provides a soft approximation of the optimal prediction-to-ground-truth assignment. DHN allows estimating the correspondences between object tracks and ground truth objects to compute differentiable proxies of MOTA and MOTP, which are in turn used to optimize deep trackers directly.

2.3.2 Joint Detection and Re-Identification Association

Most modern multiple object tracking algorithms adopt multiple cues association to tackle the data association problem. Appearance descriptors or appearance representations are robust cues due to their non-spatial constraints for camera motion scenes. When bounding box association engaging assignment problems, the appearance feature is an auxiliary means to make association decisions.

Traditional methods use hand-crafted descriptors for the appearance model, such as color histograms and optical flow-based descriptors as association cues. Under this condition, data association can be posed as a maximum-a-posteriori (MAP) estimation problem by seeking an optimal set of tracks as a conditional distribution of sequential track states. Several methods perform inference using conditional random fields (CRFs) [59], Markov chain monte Carlo (MCMC) [60] or a variational expectation-maximization [61].

After deep learning shows its power at appearance feature representation, learning-based appearance representation is often employed to improve data association. The most representative online data association algorithm is **DeepSORT** [62]. A standard Kalman filter with constant velocity motion and linear observation model propagates a target's identity into the next frame. DeepSORT trains a CNN on a large-scale person re-identification dataset to make it well suited for person appearance description. Mahalanobis distance,

instead of IoU, between predicted Kalman states and newly arrived detection is used to incorporate the cosine distance between tracks and detection to define the cost matrix. The solver of the cost matrix is Hungarian algorithm.

Tracker++ [63] convert a detector into a tracker by exploiting the bounding box regression of the detection network to predict the position of an object in the next arriving frame. Tracker++ extended the detection network with two extensions. **i**. a re-identification siamese network to discriminate and re-identify objects by appearance, particularly for the crowded scene association problem. **ii**. apply a camera motion compensation (CMC) for moving cameras by aligning frames via image registration using the enhanced correlation coefficient maximization [64].

2.3.3 Metrics

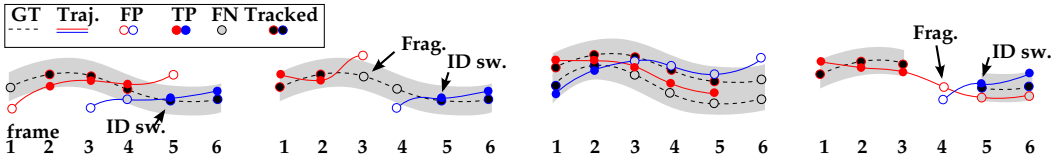


Figure 2.4 Four cases illustrating tracker-to-target assignments [2].

To provide common experimental setup for fairly test and compare different trackers, a group of standard metrics has been established and used in almost every multiple object tracker. Fig. 2.4 illustrates some basic definitions.

- **FP**: The total number of false positives.
- **FN**: The total number of false negatives (missed targets).
- **Frag**: The total number of times a trajectory is fragmented (i.e. interrupted during tracking).
- **MT** *(Mostly Tracked trajectories)*: Mostly tracked targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span.
- **ML** *(Mostly Lost trajectories)*: The ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span.
- **FAF**: The average number of false alarms per frame.
- **IDF1**: ID F1 Score [65]. The ratio of correctly identified detections over the average number of ground-truth and computed detections.
- **IDSW**: The total number of identity switches (ID switch ratio = ID switches / recall)[66]

— **MOTA** *Multiple Object Tracking Accuracy [67]*:

$$MOTA = 1 - \frac{(FN + FP + IDSW)}{GT} \in (\infty, 1] \tag{2.1}$$

This measure combines three error sources: false positives, missed targets and identity switches. where GT is the number of ground truth boxes.

— **MOTP** *Multiple Object Tracking Precision*: the misalignment between the annotated and the predicted bounding boxes.

$$MOTP = \frac{\sum_{t,i} d_(t,i)}{\sum_t c_t} \tag{2.2}$$

where c_t denotes the number of matches in frame t, and $d_{t,i}$ is the bounding box overlap between the hypothesis i with its assigned ground truth object.

2.4 Conclusion

For MOT, object detection and data association are two essential components. Besides, appearance representation technologies are adopted in enormous modern tracking systems to improve the data association robustness. The utilization of deep convolutional neural networks prompt detection technologies and re-identification technologies to an in-depth understanding of the essence of these problems. These advancements dramatically reduce the difficulties of designing multiple object tracking algorithms. The state-of-the-art methods take object detection as the core to improve tracking. For example, FairMOT[68] redesigned CenterNet[16], and JDE[69] redesigned YOLOv3[10] for faster and more accurate pedestrian detection. As a consequence, the tracking accuracy and inference speed have been greatly improved. Improving the detection performance is always the key to improving the tracking performance.

Re-identification technology, as an auxiliary means, aims to help improve association robustness when the spatial-scale constraints encounter association confusions. For those moving camera or low framerate fast-moving scenes, re-identification features are underestimated and hardly ever used alone to make association decisions. MAT[57] and Tracker++[63] propose to use an additional ECC [64] module to align frames with compensating the camera motion. Differ from these methods, we believe it is a promising application direction to re-evaluate the importance of re-identification technologies for the robust tracking for more complicated scenes.

For the data association, learning-based methods like DHN or other non-learning-based methods like SORT, DeepSORT, and MAT are focusing on the inter-frame association.

Approaching human vision system multi-frame association technologies are not being developed due to the enormous video data redundancy. A number of existing algorithms [70, 71, 72] using recurrent neural networks, like LSTM [73], to processing multiple association cues. their methods based on de-redundant detection data are still spatial-scale association technologies.

CHAPTER 3

A Deep-Frozen Transfer Learning Framework for Person Search

3.1 Introduction

Person re-identification (Re-ID) aims to match manually cropped pedestrian images between queries and candidates in multiple camera surveillance systems. It is of great relevance in application scenarios in public security and safety fields, such as multi-target multi-camera tracking [74] and person verification [75]. In recent years, with the powerful feature representation capabilities of deep convolutional neural network (CNN), person Re-ID attracting much attention in the computer vision community has achieved impressive results on many large-scale datasets and benchmarks [76, 77, 78]. However, person Re-ID itself does not have a pedestrian detection module. Without human labor intervention, applying person Re-ID in real-world scenarios requires three sequential steps: **i**. A detection network detecting pedestrians. **ii**. A module cropping and resizing pedestrian images. **iii**. A Re-ID network extracting discriminative features. As an implementation, person search is a new research domain of integrating detection and Re-ID together and aims at locating a queried person over a gallery of raw images. It is facing more domain-specific difficulties than person Re-ID [48]. Typical such kind of approach is independent person search (IPS) [48, 45, 46, 47], in which detection and Re-ID are treated as two independent tasks. It is popular due to its easy integration of detection and Re-ID technologies and has achieved very well search accuracy. Yet, the two main drawbacks of IPS are: First, due to occlusion, background clutter, and low image resolution, pedestrian

detection results are usually unaligned and inferior to manually cropped images; this may result in the loss of critical appearance features [53, 54]. Second, detection network, image cropping and resizing, and Re-ID network heavily consuming computing resources would increase computational costs.

To overcome these shortcomings, in 2014, Xu et al. [49] proposed a combination framework consists of sliding window pedestrian detection and handcrafted feature matching. Later, a number of learning based solutions were proposed [1, 50, 51, 52, 53, 54, 55]. In 2017, Xiao et al. proposed the first CNN based framework named OIM[1] that jointly implements pedestrian detection and appearance feature extraction in an end-to-end manner and achieved a top-1 accuracy of 78.7% on CUHK-SYSU, which is a person search dataset. The success of OIM inspired the latter joint person search (JPS) solutions. Liu et al. [52] studied the loss function of OIM and proposed a multi-loss fusion strategy to speed up the network convergence and compact the intra-class distance, and their approach achieved 79.9% top-1 accuracy. LCGPS [53] introduced a relative attention module to employ the scene context cues to decrease the search confusion, and it gave a top-1 accuracy of 86.5%. QEEPS [54] using a query-guided strategy to improve the feature representation between query and gallery images has achieved the highest top-1 accuracy of 89.1%. It is worth noting that these JPS solutions are still not as accurate as IPS solutions, e.g., in [45], the top-1 accuracy is 89.9%, and in [47], the top-1 accuracy is 91.7%.

Overviewing the newly proposed JPS solutions [79], a noteworthy issue is that as a hybrid field of object detection and person Re-ID, the freshly proposed detection and Re-ID technologies have not been applied. One reason is that we have to carefully consider the technical details of a detection network such as image augmentation technologies, the design of the loss function, and the network training experience. What's more, train a JPS solution has a great demand for GPU memory and need to carefully design the Re-ID part to avoid tasks mutual interference. These difficulties hinder us from developing new frameworks. OIM implemented relatively simple on Faster RCNN [13] is a starting point for many proposed JPS solutions. However, without introducing new methods, the improvements that could be made to OIM will reach a bottleneck sooner or later. This motivates us to seek for a new framework that could expediently join cutting-edge object detection and Re-ID technologies. Other than the idea of designing an IPS network, we propose a counter-intuitive framework which uses transfer learning training a Re-ID network for extracting discriminative features from the sharing data of the detection network while keeping the entire detection network frozen; this framework is named deep-frozen transfer learning framework. One superiority of this framework is that, in facing the increasingly complex and rapid updating detection networks, we could save the time of studying the details of detectors. Meanwhile, by sharing data, the Re-ID network reusing feature of the detection network could be designed lightweight. Yet, intuitively,

one imperfection of this framework is its high demand for the detection network to retain discriminative information. MGTS [48] argues that object detection and Re-ID are two contradictory tasks. Object detection treats all people as one class. The representations focus on the commonness, i.e., the shape and posture of different people. Whereas in person Re-ID, different people are deemed as different classes. The goal is to maximize the differences in people's appearance, i.e., hair, clothes, and shoes in between. We agree that this task contradiction is a reason for the low performance of JPS. As a supplement, we would like to point out that modern detectors can perform inter-class classification and can segment intra-class instances [80, 14]. This means detection networks implicitly contain discriminative information; one evidence is the success of OIM. Our proposed framework freezing the detection network would not degrade the detection performance. In order to validate the Re-ID network, a challenging task of this chapter is to verify the following two critical arguments:

— Detection networks contain sufficient discriminative features, i.e., color and detailed texture, for Re-ID.

— The shared feature of detection networks is highly compatible with the task of Re-ID that can be used for Re-ID with minimum conversion.

To this end, we propose to employ pre-trained YOLOv3[10] as underlying detector. Then, by using a channel attention mechanism, an adaptive transfer learning network (ATLnet) is proposed to convert the multi-scale fused detection-purpose feature map to a Re-ID-purpose feature map. Last, inspired by MGN[40], we design a multi-branch Re-ID network called multiple descriptor network (MDnet) to extract multi-part feature descriptors. The design of the loss function is inspired by OIM, in which a non-parametric online instance matching loss is used to tackle the training difficulties. To further illustrate the effectiveness of the proposed solution, besides YOLOv3, we also evaluate some mainstream detection networks, i.e., YOLOv4 [23], Mask RCNN [14], and CenterNet [16].

The contributions of this chapter include the following three aspects.

— Using deep-frozen transfer learning, the proposed joint person search framework could easily integrate state-of-the-art detection networks and do not need to concern the implementation details. This helps to reduce design difficulties.

— The proposed ATLnet and MDnet combine multi-scale, multi-part discriminative feature representation to improve the Re-ID accuracy.

— We use multiple mainstream detectors to design experiments and achieved accuracy far superior to other methods on CUHK-SYSU [1] and PRW [81] datasets. The experimental results prove an important conclusion; that is, the feature representation of detection networks are highly compatible with Re-ID. The Re-ID network training is not from scratch but starting with these high compatible features. A shallow-designed Re-ID network would work well.

3.2 Related Works

3.2.1 Transfer Learning

In computer vision tasks, transfer learning consists of reusing extracted fixed features from ImageNet pre-trained networks for new tasks [82]. It plays a vital role in many vision tasks such as object detection [83], action recognition [84], and image segmentation [85]. Before starting a specific task, backbone network like VGG[18], ResNet[17] and Darknet [10] will firstly train on ImageNet [19] for learning general-purpose features. There are three strategies of sharing parameters (layers) [86] for transfer learning:

- **Deep-Training**: Refine the entire backbone network with new tasks. Most modern vision tasks employ this strategy [10, 23, 14, 13, 16, 83, 84, 85].
- **Shallow-Training**: Refine a part of the parameters of the backbone network with new tasks.
- **Deep-Frozen**: Freeze the entire backbone network, which will serve as a data provider [51, 53].

In order to keep the training independence of the detection network, the third strategy, **Deep-Frozen** is adopted in our framework.

3.2.2 Pedestrian detection

Object detection solutions—YOLO [10, 23] and RCNN [13, 14] are currently the most commonly used detectors. One-step detectors like YOLOv3 or YOLOv4 achieving a good balance between speed and accuracy and truly meeting real-time requirements are the first considered detectors when designing a real-world application. RCNN series detectors, including Fast RCNN [12], Faster RCNN [13], and Mask RCNN [14], famous for their high accuracy, are the most typical two-step detectors and are applied in a variety of research topics. Modern detectors like YOLOv3, YOLOv4, and Mask RCNN equipped with FPN for dealing with the multi-scale problem are selected in this chapter as underlying detection networks. In addition to these detectors, anchor-free detectors [16, 15] also begin to attract attention. CenterNet [16] is also selected.

3.2.3 Person Re-identification

Person Re-ID driven by deep CNN is undergoing explosive development in recent years. There are some significant achievements made to person Re-ID concerning the following three aspects. The first type of work uses attention mechanisms for multi-scale, multi-part

feature representation [40, 87, 38, 44, 88]. MGN [40] propose to use a multi-branch deep CNN architecture for global and local feature representations. OSnet [44] propose to use depthwise separable convolutional layers to reduce the network parameters; meanwhile, a multiple receptive scale bottleneck was proposed for learning multi-scale features. The introduction of multi-part, multi-scale concepts achieved remarkable improvements to the accuracy of person Re-ID. The second type of work mainly uses a generative adversarial network to augment training data [89, 90]. In [90], DGnet employed a generative model to exchange appearance and structure codes between pedestrian images for generating new training samples. It explicitly stated that the color, texture, and style of pedestrians' hair, clothes, and shoes are the key components that constitute the pedestrian appearance, the pedestrian posture excluded. The third type of work uses the prior spatial-temporal constraint to narrow the search scope[91].

3.2.4 Person Search

IPS solutions, including [45, 46, 47], taking the multi-scale problem as the key to improving Re-ID accuracy, have achieved better accuracy than JPS solutions. MGTS [48] integrating pedestrian detection and pedestrian segmentation make an effort to tackle the background and foreground (pedestrian) cues separately. JPS solutions were discussed in section. 3.1 that most JPS solutions focus on improving the performance of OIM and are short of introducing freshly proposed detection and Re-ID technologies. As for the training datasets, CUHK-SYSU [1] and PRW [81] are commonly used in recent person search solutions. Our experiment is also carried out on these two datasets.

3.3 Proposed Global Network Architecture

3.3.1 Detection Subnetwork

Modern detection network generally consists of three parts, including a backbone network pre-trained on ImageNet [19], a neck (feature pyramid network) which collects feature maps from multiple backbone stages for multi-scale feature representation, and a head network which used for object classification and bounding box regression; see Fig. 3.1.

During training, the inputs to the network are images, ground truth bounding boxes, and labels. Each image corresponds to a group of labeled bounding boxes, indicating the *id* and regions of interest (ROIs). During network inference, for query images, the input to the network are images and manually specified query bounding boxes, and for gallery

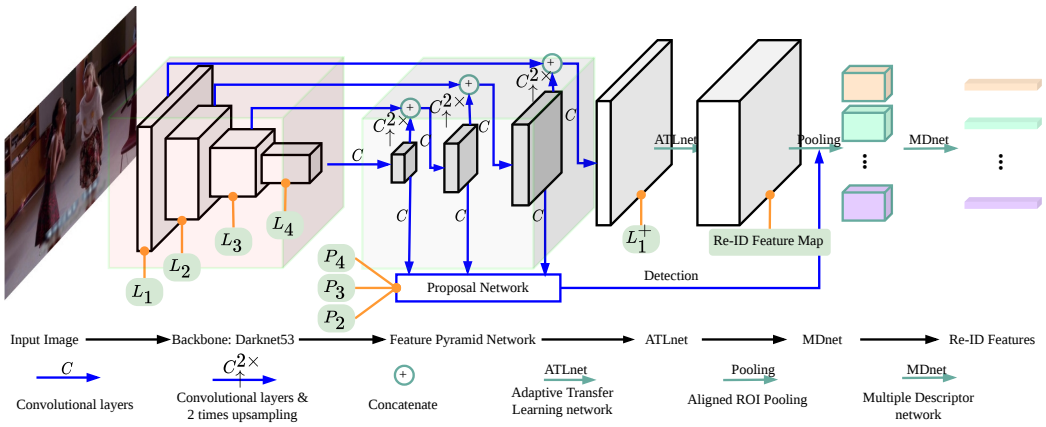


Figure 3.1 Global network architecture: An image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ compute through backbone network (darknet53). Feature pyramid network (FPN) connect the backbone and fuse multi-scale feature maps. Adaptive transfer learning network (ATLnet) transfer the fused detection-purpose feature map to Re-ID-purpose feature map. Then, an aligned ROI pooling layer pool feature regions from the Re-ID feature map. Last, the feature region fed into the multiple descriptor network (MDnet) to get multi-part Re-ID feature vectors.

images, the input is only images.

Backbone serves as a generate-purpose feature extractor undertaking a great majority of the network calculation work. The backbone of YOLOv3 is darknet53 which consists of 52 convolution layers, the output tensor of the $9^{th}, 26^{th}, 43^{th}$, and the 52^{th} convolution layers are $4\times, 8\times, 16\times$ and $32\times$ downsampled feature maps with respect to the input image size and are marked as L_1, L_2, L_3, and L_4.

Feature pyramid network (FPN) [22, 92, 93] connecting backbone and proposal network is usually used for a detection network dealing with the multi-scale problem. The FPN of YOLOv3 receives feature maps from the backbone. From top to bottom ($L_4 \rightarrow L_2$), the feature maps are upsampled and fused down. For example, the feature map L_4 firstly processed through:

$$\begin{aligned} L_4^{'} &= conv^{(5)}(L_4) \\ P_4 &= conv^{(2)}(L_4^{'}) \end{aligned} \tag{3.1}$$

where $conv^{(i)}$ is a block of i convolution layers. P_4 is a proposal map used for object detection. The feature map $L_4^{'}$ is upsampled and fused with L_3 by:

$$\begin{aligned} L_4^{''} &= up^{2\times}(conv^{(1)}(L_4^{'})) \\ L_3^{+} &= concat(L_4^{''}, L_3) \end{aligned} \tag{3.2}$$

where $up^{2\times}$ is a $2\times$ upsampling layer, *concat* is an operation that concatenates the input tensors in channel dimension. Repeating the process of (3.1) and (3.2) (L_3^+ instead of L_4), the FPN generates three proposal maps ($P4 \rightarrow P2$) for object detection and an $8\times$ downsampled feature map $L_2^{'}$. We further fuse $L_2^{'}$ and L_1 by use the pipeline of (3.2); the final fusion L_1^+ is a $4\times$ downsampled feature map. We design the FPN mainly considering the following two points:

— High-resolution feature map has better detailed feature representation is critically important for small-size targets. The fused feature maps aggregating large-scale abstract features and small-scale detailed features are well-adapted for Re-ID feature extraction.

— The FPN of YOLOv3 is reused for maximizing computational efficiency. Those detectors that do not have FPN, purely use the pipeline (3.2) would also work.

Proposal network is the head of the detection network that receives feature maps ($P4 \rightarrow P2$) from FPN and outputs bounding boxes and classes information. The original implementation of YOLOv3 trained on COCO [94] has multiple categories of labels. Our proposed network only consider the label '*person*'.

3.3.2 The Proposed ATLnet

One main contribution of this work is the design of a Re-ID network constructed with an adaptive transfer learning network (ATLnet) and a multiple descriptor network (MDnet).

Object detection and person Re-ID are two contradictory tasks [48]; the tendencies of both networks towards extracting features are well distinct. Assuming that the backbone of the detection network retains discriminative information for Re-ID, the mission of the ATLnet is to reorganize features according to their importance. We intend to enhance those features that are discriminative for Re-ID while weakening those features that are less important. We adopt a squeeze-and-excitation block [95] to adaptively re-weight the feature maps channel-wise. The ATLnet could be written as:

$$ATLnet(L_1^+) = conv_{dw}^{(1)}(L_1^+ f_c^s(f_c^r(gap(L_1^+)))) \tag{3.3}$$

The processes of the ATLnet is as follow:

— A tensor $L_1^+ \in \mathbb{R}^{H_i \times W_i \times C_i}$ received from FPN is firstly pooled to be a $1 \times 1 \times C_i$ vector by a global average pooling layer *gap*.

— A fully connected layer f_c^r with *relu* activation down-samples the vector's dimension to C_i/τ, where τ is a reduction ratio

— The second fully connected layer f_c^s with *sigmoid* activation restores the vector's dimension to C_i and multiplies L_1^+ channel-wise.

— A depthwise separable convolutional layer $conv_{dw}^{(1)}$ [44] mounted to adjust the channel of the tensor and outputs a Re-ID feature map.

This simple conversion is sufficient to convert the detection-purpose feature map to a Re-ID-purpose feature map.

3.3.3 The Proposed MDnet

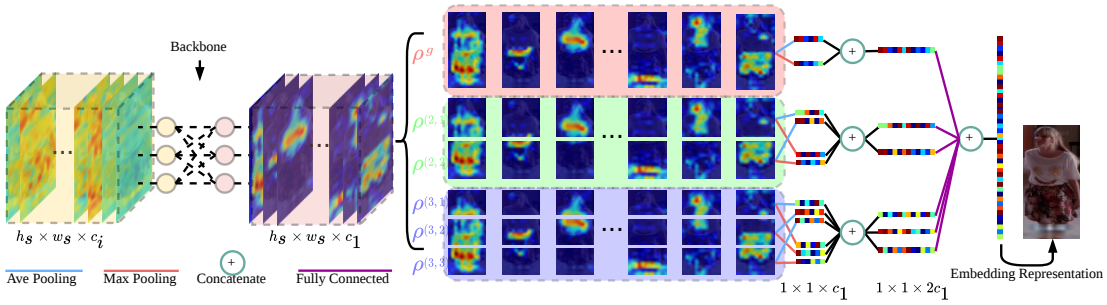


Figure 3.2 The architecture of MDnet, this figure shows an inference process of Re-ID feature representation.

After getting the Re-ID feature map, an aligned ROI pooling layer [14] with $h_s \times w_s$ sampling template bi-linearly samples the region of interest from it. In principle, a smaller sampling template could give higher computational efficiency but is prone to under-sampling. We intend to use a high-resolution feature map for keeping more detailed features; a large sampling template, i.e., 24×12, is required.

A feature region received from aligned ROI pooling layer has shape $h_s \times w_s \times c_i$. The

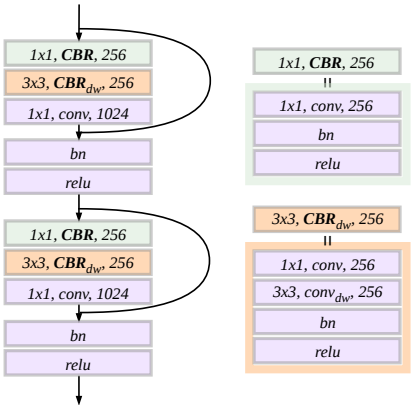


Figure 3.3 The backbone of MDnet consists of two residual units (left). The convolutional layers with 3×3 kernel are replaced by depthwise separable convolutional layer. The subscript '$_{dw}$' means depthwise.

backbone of the MDnet is shallow-designed that consists of two residual units to learn attention to the feature region; see Fig. 3.3. We apply depthwise separable convolutional layers [44] in it to reduce the network parameters. It worth noting that we keep the resolution of the feature region to preserve small-scale features, the output tensor has shape $h_s \times w_s \times c_1$. The tensor could be written as $\vartheta = \{\rho_i\}_{i=1}^{c_1}$, where $\rho_i \in \mathbb{R}^{h_s \times w_s}$ is the i^{th} channel of the feature region. The pixel in ρ_i expresses the intensity of a specific abstract feature at the corresponding location. In this chapter, the Re-ID feature representation is defined as the description of the feature region ϑ that is written as $\Re = \{\chi(\rho_i)\}_{i=1}^{C_1}$, where $\chi(\rho_i)$ is the description of ρ_i that consists of one or more one-dimensional descriptors. Modern person Re-ID algorithms usually adopt max-pooling to get descriptors, that is, $\Re = \{\rho_i^{max}\}_{i=1}^{C_1}$, where ρ_i^{max} is the maximum value of ρ_i. In our solution, the idea of designing the Re-ID network is to maximize the information usage of the feature region ϑ. First, ρ_i is divided into multiple parts: $P_i = \{\rho_i^g, \rho_i^{(2,1)}, \rho_i^{(2,2)}, \rho_i^{(3,1)}, \rho_i^{(3,2)}, \rho_i^{(3,3)}\}$. Where $\rho_i^g = \rho_i$ is the global branch, $\rho_i^{(2,j)}$ is the j^{th} part of the two-split branch, and $\rho_i^{(3,j)}$ is the j^{th} part of the three-split branch. Then, we apply both ave-pooling and max-pooling to P_i, the description of ϑ is written as: $\Re = \{(P_i^{max}, P_i^{ave})\}_{i=1}^{C_1}$; each description $\chi(\rho_i)$ consisting 12 descriptors is relatively more complete.

The architecture of MDnet is shown in Fig. 3.2. The overall processes are as follow:

— The input tensor is firstly fed into the backbone to get the attention feature region ϑ.

— Applying multi-branch architecture, the tensor ϑ is divided into multiple parts. The local branch includes a two-split brach and a three-split branch that horizontally divide the ϑ into 2 and 3 equal-size sub-tensors.

— Incorporating the global branch, the 6 sub-tensors are pooled by a global max-pooling layer and a global ave-pooling layer. The pooled vectors are pairwise concatenated and separately fed into 6 fully connected layers for dimension reduction.

— Concatenate the vectors to get the final representation of appearance embedding.

In the inference phase, the Re-ID feature vector is $l2$ normalized to measure cosine similarity. In the training phase, the vector is sent to the *loss* module to get losses.

3.3.4 Online Instance Matching and Loss

Unlike person Re-ID, training a person search network needs to input a full image rather than a small cropped pedestrian image. Limited by the image size and the network complexity, it is usually computationally expensive to pile enough images in a batch; therefore, it is hard to carry out effective training. The design of the loss function is inspired by OIM [1]. Supposing there are N identified pedestrians, a data structure called

look up table (LUT): $\mathbf{L} \in \mathbb{R}^{N \times D}$ sorted by id is used to keep all identities, where D is the dimension of feature vectors. Given a feature vector $v_i \in \mathbb{R}^D$ where i is the id, there is one and only one positive sample in the LUT, which is denoted as $\mathbf{L}_i$ (the i^{th} item of $\mathbf{L}$). The rest items of the LUT are all negative with respect to v_i. The cosine similarity between v_i and $\mathbf{L}_j$ is:

$$s(v_i, \mathbf{L}_j) = \frac{v_i^T \mathbf{L}_j}{\|v_i\| \, \|\mathbf{L}_j\|} \tag{3.4}$$

where $\|\cdot\|$ is $l2$ norm. The range of cosine similarity is between $[-1, 1]$, -1 means the two vectors are completely different, and 1 means the two vectors are completely the same. Thus, the objective function could be written as:

$$s^{obj}(v_i, \mathbf{L}_j) = \begin{cases} 1 & \text{if } j = i \\ -1 & \text{if } j \neq i \end{cases} \tag{3.5}$$

Given M labeled targets $\mathbf{V} \in \mathbb{R}^{M \times D}$, the similarity is: $S_{ij} = s(\mathbf{V}_i, \mathbf{L}_j)$, where S_{ij} is the i^{th} row and j^{th} column element of the similarity matrix $S \in \mathbb{R}^{M \times N}$. The corresponding objective matrix is $S^{obj} \in \mathbb{R}^{M \times N}$. Due to the extreme imbalance between positive and negative samples, using binary cross-entropy will make the network unable to converge. So, we firstly perform a softmax on each row of the similarity matrix:

$$\tilde{S}_i = \frac{exp(S_i t)}{\sum_{j=1}^{N} exp(S_{ij} t)} \tag{3.6}$$

where S_i means the i^{th} row of S and t ($1/t$ is called temperature parameter) is used to control the sensitivity of the loss with respect to the similarity. Larger t results in softer changes. The new objective matrix is:

$$\tilde{S}_{ij}^{obj} = \begin{cases} 1 & \text{if } S_{ij}^{obj} = 1 \\ 0 & \text{if } S_{ij}^{obj} = -1 \end{cases} \tag{3.7}$$

where $\tilde{S}_{ij}^{obj}$ is the i^{th} row and j^{th} column element of the objective matrix $\tilde{S}^{obj}$. The loss of the network could be written as:

$$loss = -\frac{1}{M} \sum_{i=1,j=1}^{M,N} log(\tilde{S}_{ij} \tilde{S}_{ij}^{obj}) \tag{3.8}$$

At the end of each batch iteration, update the LUT by:

$$\mathbf{L}_n \leftarrow \gamma \mathbf{L}_n + (1 - \gamma) v_n \tag{3.9}$$

where $\gamma \in [0, 1]$.

3.4 Experiments and Evaluation

In this section, we conduct experiments to validate the proposed framework. In 3.4.1 and 3.4.2, we introduce the implementation details and the datasets information. In 3.4.3, we compare our proposed networks with some state-of-the-art person search solutions. In 3.4.4, we extended evaluate other mainstream detectors and visualize the experimental results to verify the effectiveness of the proposed framework. In 3.4.5, we conduct multiple ablation experiments to verify the parameters selection problem. In 3.4.6, we evaluate the running speed of our proposed networks.

3.4.1 Implementation details

Our solution is implemented on Keras with the TensorFlow backend. For the network training, we use a Geforce GTX 1080Ti graphics card with 11G memory. We add $l2$ regularizers to the trainable parameters of the convolution and fully connected layers with value 0.0005. The reduction ratio τ in ATLnet is set to 16. The sampling template of the aligned ROI pooling layer is set to 24×12. The parameter t in equation (3.6) is set to 15. The update rate of γ of LUT is set to 0.5. We use a batch size of 6. In each iteration batch, we select 5000 negative samples for equation (3.6) and equation (3.8). We define an epoch as a traversal of the entire training set. The learning rate is set to 0.01 in the first 20 epochs and then the learning rate decreases to 0.001 for the remaining 20 epochs. During training, we freeze the entire detection network.

3.4.2 Datasets

CUHK-SYSU and PRW are used to train and evaluate our proposed networks. The images of CUHK-SYSU were captured from street snaps using hand-held cameras or obtained from some movie snapshots. It includes 18184 images with 5532 labeled identities for training and 2900 labeled identities for evaluation. The PRW dataset was captured from a University campus; it consists of 11816 images from 6 fixed cameras with 482 labeled identities for training and 450 labeled identities for evaluation. CUHK-SYSU and PRW correspond to different application scenarios. We show statistics of the relative size of the training targets for the two datasets in Fig. 3.4. We define the target relative size as:

$$s = \frac{max(w_b, h_b)}{max(w_I, h_I)} \tag{3.10}$$

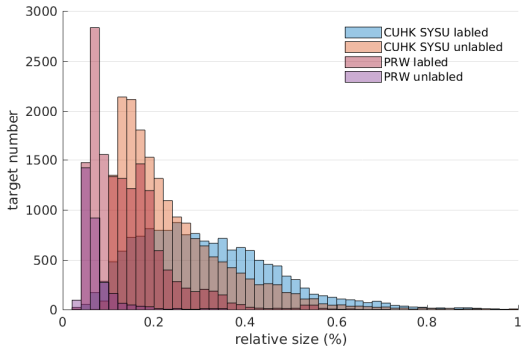


Figure 3.4 Targets relative size distribution

where w_b and h_b are the width and height of the bounding box b, w_I and h_I are the width and height of the image I which contains b.

Fig. 3.4 shows that CUHK-SYSU and PRW are well distinct in the target size distribution. To validate our proposed framework, we train our network on these two datasets conjunctively. The size of LUT is 6014 that consists of 5532 training identities from CUHK-SYSU, and 482 training identities from PRW. Each epoch contains 15420 training images, 9716 from CUHK-SYSU and 5704 from PRW. To prevent over-fitting, we randomly flip the image horizontally and randomly adjust the image brightness between $[0.7, 1.3]$. We also add random noise (0.1 times the bounding box size) to the ground truth bounding boxes.

3.4.3 Evaluation Results

Table 3.1 Evaluation comparison results for YOLOv3 on CUHK-SYSU dataset

Methods	Years	CUHK-SYSU (50)		CUHK-SYSU (100)		Input Size
		mAP (%)	top-1 (%)	mAP (%)	top-1 (%)	
OIM[1]	2017	81.9	82.6	78.0	78.7	600 × 600
LCGPS[53]	2019	87.8	-	84.1	86.5	720 × 576
QEEPS[54]	2019	-	-	88.9	89.1	900 × 900
MGTS[48]	2018	84.8	-	83.0	83.7	-
CLSA[46]	2018	-	-	87.2	88.5	-
DHFF[47]	2019	-	-	90.2	91.7	-
YOLOv3-MDnet	-	90.5	91.8	89.2	90.6	608 × 608
YOLOv3-MDnet^{+}	-	**94.1**	**95.3**	**93.2**	**94.5**	608 × 608
YOLOv3-MDnet$^{+}_{+}$	-	92.2	92.9	91.4	92.0	608 × 608

Table 3.2 Evaluation comparison results for YOLOv3 on PRW dataset

Methods	Years	PRW		Input Size
		mAP(%)	top-1(%)	
OIM[1]	2017	36.9	75.7	900 × 900
LCGPS[53]	2019	33.4	73.6	720 × 576
QEEPS[54]	2019	37.1	76.6	900 × 900
MGTS[48]	2018	32.6	72.1	-
CLSA[46]	2018	38.7	65.0	-
DHFF[47]	2019	41.1	70.1	-
YOLOv3-MDnet	-	43.4	83.0	608 × 608
YOLOv3-MDnet^{+}	-	52.0	86.0	608 × 608
YOLOv3-MDnet$^{+}_{+}$	-	**54.9**	**87.2**	608 × 608

Table 3.3 Evaluation comparison results for other mainstream detectors on CUHK-SYSU

Methods	CUHK-SYSU (50)		CUHK-SYSU (100)		Input Size
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)	
YOLOv4-MDnet	91.6	92.1	89.7	90.2	608 × 608
MRCNN-MDnet	92.7	93.8	91.6	92.7	1024 × 1024
YOLOv4-MDnet^{+}	**94.6**	**95.2**	**93.7**	**94.5**	608 × 608
CenterNet-MDnet^{+}	93.9	95.0	92.9	94.0	608 × 1088

Table 3.4 Evaluation comparison results for other mainstream detectors on PRW dataset

Methods	PRW		Input Size
	mAP(%)	top-1(%)	
YOLOv4-MDnet	41.2	80.7	608 × 608
MRCNN-MDnet	47.4	83.3	1024 × 1024
YOLOv4-MDnet^{+}	**55.1**	**86.3**	608 × 608
CenterNet-MDnet^{+}	54.6	86.3	608 × 1088

Two commonly used metrics—mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC top-K) [1] are used for evaluation. We design three approaches that employ YOLOv3 as underlying detector:

— **YOLOv3-MDnet**, in which YOLOv3 pre-trained on COCO [94] dataset.

— **YOLOv3-MDnet^{+}**, in which YOLOv3 is firstly retrained on CUHK-SYSU and PRW dataset for pedestrian detection, and then train the Re-ID module while keeping the detector frozen.

— **YOLOv3-MDnet$^{+}_{+}$**, in which we use the well-trained YOLOv3-MDnet^{+} as a starting point and retrain the entire network for Re-ID. This network serves as a pure Re-ID network and the detection is getting from YOLOv3-MDnet^{+}.

The evaluation comparison results are shown in Table. 3.1 and Table. 3.2. The two most recent JPS solutions LCGPS [53], QEEPS [54] and the baseline network OIM [1] are included. Three IPS solutions MGTS [48], CLSA [46], and DHFF [47] are also included in the table. Based on these experiments, we could note the following peculiarities of the proposed solution:

(i) Our proposed framework effectively works with YOLOv3 backbone. The accuracy of YOLOv3-MDnet close to the state-of-the-art JPS and IPS solutions. QEEPS achieved 88.9% mAP and 89.1% top-1 accuracy on CUHK-SYSU (100) (where 100 is the gallery size), the accuracy is lower than YOLOv3-MDnet, which has an accuracy of 89.2% mAP and 90.6% top-1. DHFF [47] achieved 1.0% mAP and 1.1% top-1 higher accuracy than YOLOv3-MDnet on CUHK-SYSU (100), but the accuracy on PRW is much worse than YOLOv3-MDnet, which has 2.3% mAP and 12.9% top-1 higher accuracy. YOLOv3-MDnet achieving state-of-the-art search accuracy on both the two person search datasets proved that YOLOv3 retains effective discriminative information and our proposed deep-frozen transfer learning framework effectively extracts the Re-ID features out. To further study the source of improvements, we evaluate a combination solution in Table. 3.5 named OIM+YOLOv3-MDnet in which OIM serves as a detector and YOLOv3-MDnet serves as a Re-ID network. The accuracy of OIM+YOLOv3-MDnet is close to YOLOv3-MDnet and has a significant improvement compared to OIM; this proved that the performance improvements of YOLOv3-MDnet mainly come from the Re-ID part instead of the pedestrian detection part.

(ii) The better the detection network, the better the person search accuracy. In Table. 3.1, the detection part of YOLOv3-MDnet^{+} is retrained on CUHK-SYSU and PRW. The accuracy has been significantly improved—4.0% mAP and 3.9% top-1 higher accuracy compared with YOLOv3-MDnet. In Table. 3.5, we evaluated a combination solution YOLOv3-MDnet^{+}+YOLOv3-MDnet, in which YOLOv3-MDnet^{+} serves as a detector and YOLOv3-MDnet serves as a Re-ID network. Compared with YOLOv3-MDnet, the accuracy improved by 2.2% mAP and 1.8% top-1. We also evaluated another com-

Table 3.5 Evaluation on CUHK-SYSU (100) for different combinations

Methods	mAP (%)	top-1 (%)
OIM+YOLOv3-MDnet	88.8	90.2
YOLOv3-MDnet$^+$+YOLOv3-MDnet	91.4	92.4
YOLOv3-MDnet+YOLOv3-MDnet$^+$	90.4	91.7

bination solution, YOLOv3-MDnet+YOLOv3-MDnet$^+$, in which YOLOv3-MDnet serves as a detector, YOLOv3-MDnet$^+$ serves as a Re-ID network. Compared with YOLOv3-MDnet, the accuracy improved by 1.2% mAP and 1.1% top-1. This shows that a well-trained detection network not only provides better detection but also better discriminative information.

(iii) End-to-end training YOLOv3-MDnet$^+$ is prone to overfitting. In Table. 3.1, YOLOv3-MDnet$^+_+$, using YOLOv3-MDnet$^+$ as a training starting point, is an end-to-end trained Re-ID network. It could be seen that compared with YOLOv3-MDnet$^+$, the accuracy on CUHK-SYSU decreased more than 1% while on PRW, the accuracy improved by 2.9% mAP and 1.2% top-1. This proved that when increasing the trainable parameters, the network is prone to overfitting. It must be pointed out that this argument only holds for our specific experiment setup. We believe end-to-end training would further improve the Re-ID performance but needs a meticulous design to overcome the overfitting problem.

3.4.4 Extended Experiments

To verify the universality, we selected several mainstream detectors to perform our framework. The evaluation results are shown in Table. 3.3 and Table. 3.4, in which the detection network of YOLOv4-MDnet and MRCNN-MDnet are pre-trained on COCO[94]. The detection network of YOLOv4-MDnet$^+$ and CenterNet-MDnet$^+$ are retrained on CUHK-SYSU and PRW. It could be seen that our proposed framework is effective for the listed detection networks.

To fully evaluate the performance on CUHK-SYSU, Fig. 3.5 shows the search accuracy with different gallery sizes–[50, 100, 500, 1000, 2000, 4000]. We collect the evaluation results of OIM, LGCPS, FPSP [45], CLSA [46], DHFF [47], and MGTS [48] from the corresponding papers; among them, FPSP, CLSA, DHFF, MGTS are IPS solutions. It could be seen that overall the listed methods, the accuracy decreases rapidly as the gallery size increases. Our proposed solution YOLOv3-MDnet$^+$ outperforms all other IPS and JPS solutions on all gallery sizes by a large margin.

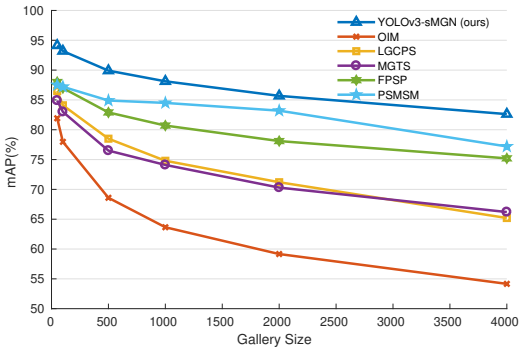


Figure 3.5 The impact of gallery size on performance

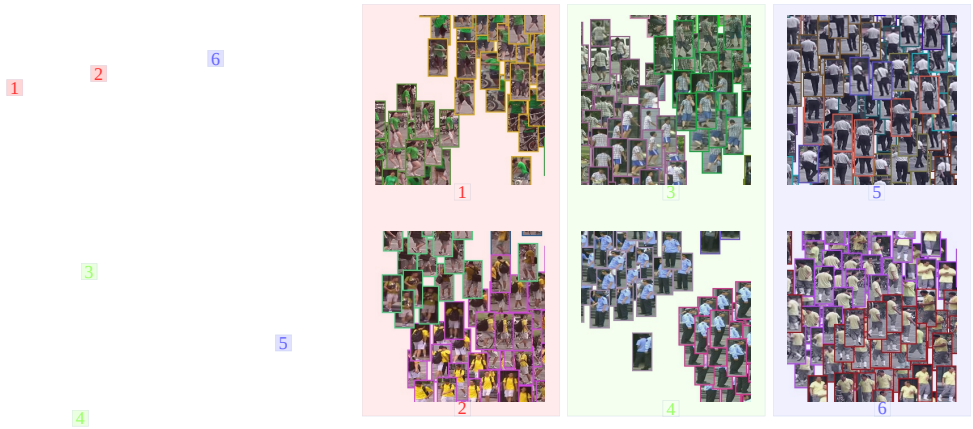


Figure 3.6 t-SNE visualization on PRW dataset. The left point cloud map includes 450 identities with 16344 points (best viewed with color and zoom in). Different color corresponds to different pedestrian identities. We select 6 regions out for analyzing.

It worth noting that all person search solutions, including ours, have worse mAP accuracy on the PRW dataset. To elaborate on this problem and further validate the effectiveness of our proposal, we visualized the search results using t-SNE [96], see Fig. 7.5. We project all the Re-ID features of the identifiable pedestrians of PRW to a 2D plane. There are 450 pedestrians consisting of 16344 pedestrian images. Each pedestrian image corresponds to a point; the points with same color indicate the same pedestrian. It could be seen that our solution has pretty good clustering effect. We select 6 region out from the point cloud map and draw the following three conclusions:

— Region 1, 2 are well distinct on the point cloud map. The pedestrians in regions 1, 2 wearing similar styles of clothes are well distinct by color. This means the pre-trained detection network retained color information.

— In region 3, 4, two appearance-similar pedestrians are distinguished by the details. Region 3, the handheld item and the shoes provide discriminative information.

Region 4, the security guards are well-distinguished by long-sleeve and short-sleeve uniforms.

— In region 5, 6, appearance-similar pedestrians are well clustered. But it is prone to give wrong search results.

The reason for the low mAP is that, in the PRW dataset, a query pedestrian has tens of hundreds of positive matches. In different camera viewports, the appearance of pedestrians may change dramatically. It is not difficult for the network to match a positive pedestrian with top-1 similarity. Still, it is hard to find out all the positive pedestrians since many negative matches closer than positive matches the appearance of the query pedestrian.

In order to step into this problem, we listed some failure cases that give the worst average precision (AP) in Fig. 3.7. It could be seen that most of the false matching cases are due to the occlusion in the query image. In Fig. 3.7, there are some dataset label errors, the query *id* 575, 661, 778, 784, and 856 are all wrong labeled cases (PRW assigned a wrong *id* to the query image), our network actually matched the correct target. What's more, for the query *id* 675 and 803, our network seemed matched the correct sample, but the dataset did not label them out. For the query *id* 726 and 919, the appearance of the false negative matching is very close to the query pedestrian.

The failure cases in CUHK-SYSU dataset were listed in Fig. 3.8. We performed a search on CUHK-SYSU with gallery size 6978 and achieved 80.0% mAP and 82.5% top-1 accuracy. The worst 36 failure cases are picked out. We found the query index 493, 1200, 1227, 1258, 1310, and 2525 are dataset label errors (The bounding box given by CUHK-SYSU point to a wrong region). 1625 and 1771 are fully occlusion cases.

The reasons for the failure cases could be summarized in the following five points.

— low image quality, including occlusion and low image resolution

— abrupt changing of illumination and appearance, especially in some movie snapshots

— pedestrian appearance lacks effective Re-ID information

— miss detection

— dataset label errors

3.4.5 Ablation Experiments

We study four factors that could significantly affect network performance.

i. The effect of selecting different feature maps on network accuracy. The outputs of YOLOv3 backbone marked as L_1, L_2, L_3, and L_4 are 4×, 8×, 16× and 32×

Table 3.6 Ablation study of selecting different feature maps

Feature Maps	CUHK-SYSU (100)		PRW	
	mAP (%)	top-1 (%)	mAP(%)	top-1(%)
L_1	90.8	91.9	48.0	85.0
L_2	92.5	93.5	45.4	81.3
L_3	91.1	92.1	30.6	72.8
L_1^+	93.2	94.5	**52.0**	**86.0**
L_2^+	**93.5**	**94.8**	47.1	82.2
L_3^+	91.4	92.5	30.1	71.7

Table 3.7 Ablation study of selecting different architectures of ATLnet

Methods	CUHK-SYSU (100)		PRW	
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
Conv	91.1	92.3	48.6	82.8
SE+Conv	**93.2**	**94.5**	52.0	**86.0**
CBAM+Conv	92.9	94.1	**52.2**	85.7

Table 3.8 Ablation study of selecting different number of residual uints of the MDnet backbone

Depth	CUHK-SYSU (100)		PRW	
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
0	86.5	88.0	37.2	69.5
1	90.6	92.0	49.3	82.2
2	93.2	94.5	**52.0**	86.0
3	**93.3**	**94.8**	53.1	**86.7**

Table 3.9 Ablation study of selecting different template sizes of the aligned roi pooling layer

Templates	CUHK-SYSU (100)		PRW	
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
12×6	87.5	88.0	34.2	68.5
24×12	93.2	94.5	52.0	86.0
36×18	**93.7**	**95.2**	**59.1**	**87.2**

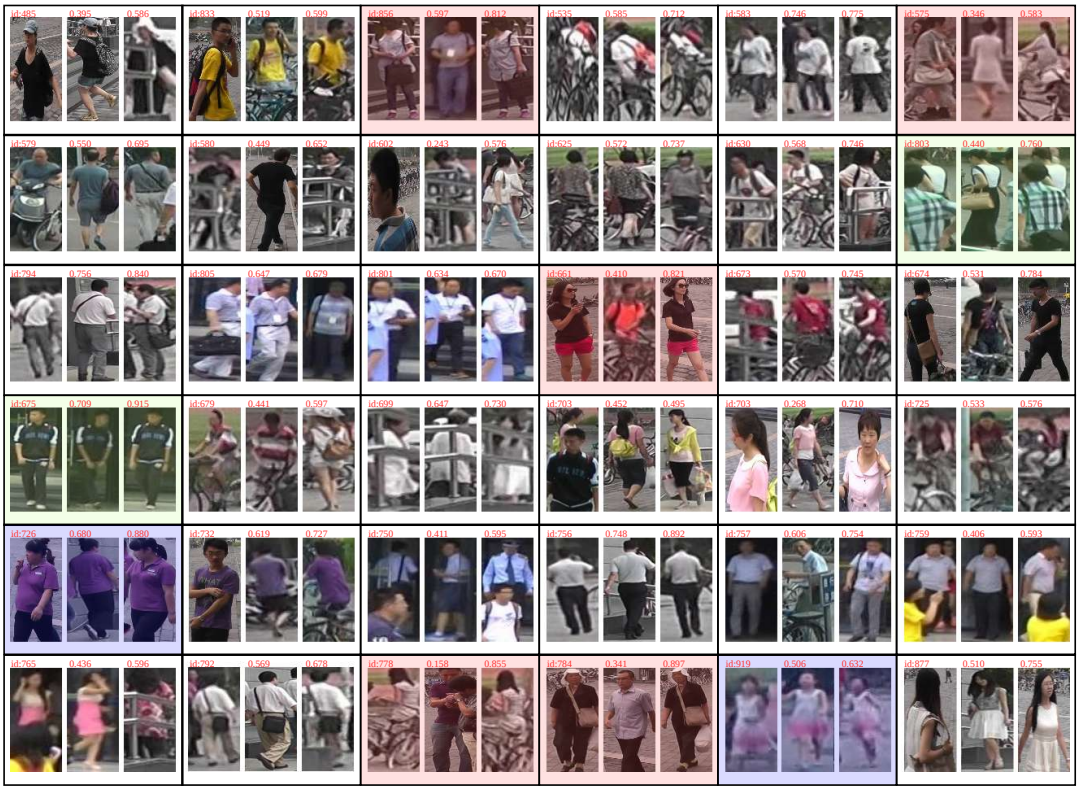


Figure 3.7 Failure cases in PRW dataset. Every 3 image is a group. The first image is a query; the number on the top is the *id* number. The second image is a false positive sample labeled by the dataset; the upper number is the similarity with respect to the query image. The third image is a false negative sample that rank top-1 the similarity. The query *id* 575, 661, 778, 784, and 856 are dataset label errors. Besides, the query *id* 803 and 675 seemed matched the correct target; the query *id* 726 and 919 have very close appearance between query images and false-negative images. For other failure cases, occlusion, low image resolution, and lack of effective discriminative texture are the main reason.

downsampled feature maps with respect to the input image size. In the FPN of YOLOv3, the feature maps are hierarchically fused and we mark the fused feature maps as L_1^+, L_2^+, L_3^+. We experiment layer by layer by selecting different feature maps among $L_1 \rightarrow L_3$ and $L_1^+ \rightarrow L_3^+$. We evaluate the trained network on CUHK-SYSU and PRW; the results are listed in Table. 3.6. It shows that the effect of selecting different feature maps are different on CUHK-SYSU and PRW. For CUHK-SYSU, when the feature map is 8× downsampled (L_2 or L_2^+), the accuracy is the highest. When selecting L_3 or L_3^+, the accuracy decrease sharply; this is due to the loss of detailed features. As for the low accuracy of selecting feature map L_1, one possible reason is undersampling due to the small sampling template (24×12) in aligned ROI pooling layer and another possible reason is L_1 lacks high-level

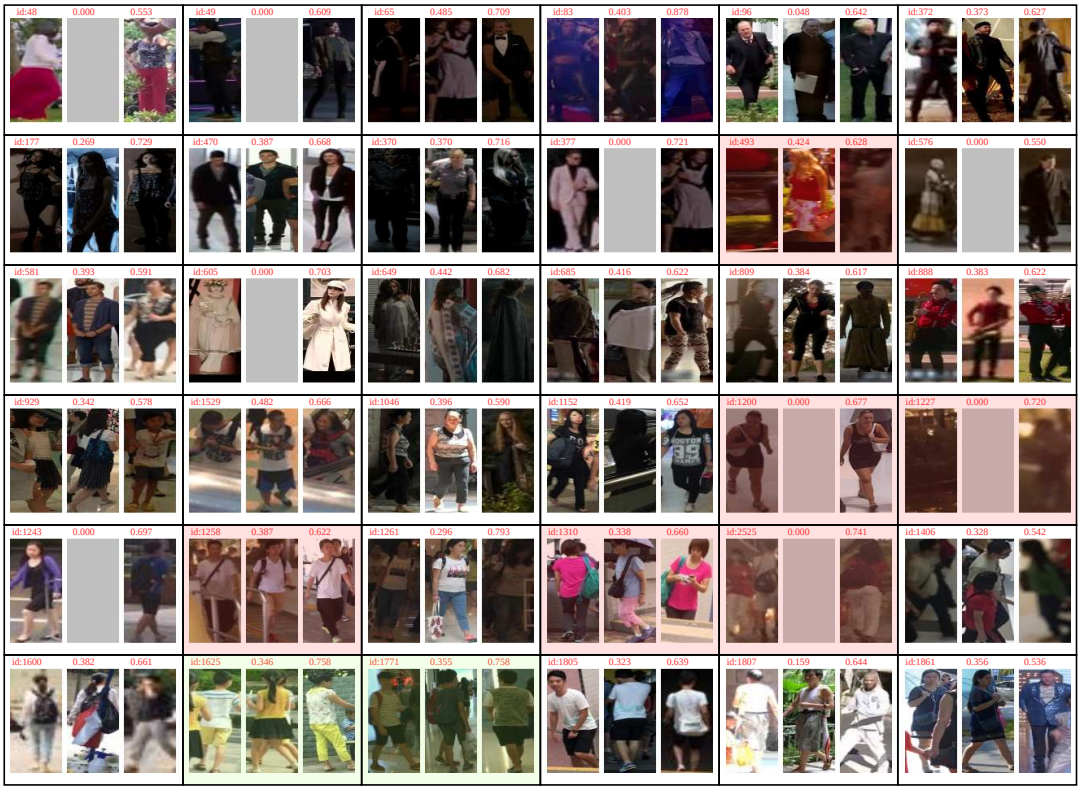


Figure 3.8 Failure cases in CUHK-SYSU dataset. Every 3 image is a group. The first image is a query and the number on the top is the query *index*. The second image is a false positive sample and the upper number is the similarity with respect to the query image. The third image is a false negative sample that rank top-1 the similarity. In some cases, the detector failed to detect a labeled pedestrian; we fill them in with a gray image. The query *index* 493, 1200, 1227, 1258, 1310, and 2525 are dataset label errors. 1625 and 1771 are fully occlusion cases. Besides, the sharp changes in brightness, background, and pedestrian posture are the main reasons for search failure.

abstract Re-ID features because it is the first stage of YOLOv3 backbone. For PRW, small targets account for a large proportion, see Fig. 3.4, the Re-ID performance no longer depends on high-level abstract features, but detailed features, that means higher resolution feature map always results in higher accuracy.

ii. **The architecture of the adaptive transfer learning network.** We tried several different combinations of solutions. **1**. pure convolutional layers. **2**. squeeze-and-excitation (SE) + convolutional layers. **3**. convolutional block attention module (CBAM)[97] + convolutional layers. We experimented with these three combinations while ensuring the same network complexity; see Table. 3.7. Compared with pure convolutional layers, SE and CBAM can significantly improve the accuracy (more than 2%).

Compared with SE, the structure of CBAM is slightly more complicated and has no obvious improvement. In our network, we use the combination of the SE + convolutional layer to design the ATLnet.

iii. **The depth of the simplified multiple granularity network.** We studied the effect of the depth of the MDnet backbone; see Table. 3.8. Increasing the depth (the number of residual units) of the backbone could slightly improve the accuracy. Decreasing the depth, i.e., zero or one residual unit, will cause a sharp drop in accuracy. Using two residual units is a trade off between accuracy and calculation efficiency.

iv. **The sampling template size of the aligned ROI pooling layer.** Large sampling template always results in higher accuracy; see Table. 3.9. Using a template size of 36×18 achieved much better search accuracy than small tempate sizes but is too computational costive. We use template size of 24×12 to achieve good balance between computing efficiency and accuracy.

3.4.6 Inference Speed

Running speed is an critical evaluation criterion for a program to be used in real-world scenarios. However, due to the influence of programming languages and underlying frameworks, it is almost impossible to compare the running speed for different proposals fairly. Purely counting the number of involved parameters could not effectively reflect the actual running speed. The size of the input images and the network architecture would also affect the speed. Instead of comparing with other solutions, we use the detectors' running speed as baseline to evaluate our networks.

To get the running time of the Re-ID module, we compare two groups of networks. **i**. pure detection network (Det), for example, YOLOv3$^+$. **ii**. detection + re-identification (Det + Re-ID), for example, YOLOv3-MDnet$^+$. The running time consists of two parts—image Preprocessing Time (PT) and network Inference Time (IT). We employed YOLOv3-MDnet$^+$, YOLOv4-MDnet$^+$, and MRCNN-MDnet to perform target detection and feature extraction for 6978 gallery images from CUHK-SYSU. The statistical running time was shown in Table.3.10.

The preprocessing time accounts for a large proportion of the total running time. When the Re-ID network is added to the detection network, the increasing of the inference time is relatively fixed—38.2 seconds for YOLOv3-MDnet$^+$ when the input image size is 608×608, 43.9 seconds for YOLOv4-MDnet, and 47.9 seconds for MRCNN-MDnet. When YOLOv3 and YOLOv4 are claimed to be realtime object detection networks, the inference time of YOLOv3-MDnet$^+$ is 343.7 seconds less than 364.3 seconds, which is

Table 3.10 Running Time for 6978 Images

Methods	Det		Det + Re-ID		Input Size
	PT(s)	IT(s)	PT(s)	IT(s)	
YOLOv3-MDnet$^+$	134.8	201.2	134.5	248.2	416×416
YOLOv3-MDnet$^+$	355.7	305.5	368.6	343.7	608×608
YOLOv3-MDnet$^+$	424.2	447.9	425.9	495.1	800×800
YOLOv4-MDnet$^+$	364.5	364.3	359.8	408.2	608×608
MRCNN-MDnet	500.7	987.2	497.6	1035.1	1024×1024

the inference time of YOLOv4. The processing time of an image is around 0.10 second. For comparison, QEEPS reported its processing speed. When the input image size is 900×1500, and the graphics card is Nvidia P6000, the image processing time is 0.3s.

3.4.7 Discussion

Our experiments proved that transfer learning from a detection network to a person search network is effective and could achieve performance beyond all other person search solutions. Although we had already given the network design and parameter selection, it is still important to explicitly explain what gives our network the performance improvement and the significance of this work contribution.

The performance improvement mainly comes from the following three points:

— The detection and Re-ID network are designed in a relatively independent manner to avoid mutual interference. Our experiments shown that the detection network keeps discriminative information for Re-ID. We can achieve the best performance for both tasks.

— Our network borrows some ideas from person Re-ID networks to realize multi-scale multi-part feature extraction. The proposed multi-scale feature fusion, ATLnet and MDnet are effective for the Re-ID feature representation.

— The state-of-the-art detectors greatly improve the accuracy of pedestrian detection.

The significance of this work is not limited to the accuracy improvement achieved on the two person search datasets. The main conclusion of this chapter is that a pre-trained detection network could preserve the discriminative information needed for Re-ID. This conclusion holds for the listed mainstream detectors. Based on this conclusion, the design of JPS solutions could follow this pipeline:

1 Train a pedestrian detector by using a state-of-the-art detection network or any other types of detectors.

2 Fuse multi-scale feature maps of the backbone of the detection network by use the pipeline of equation (3.2).
3 Train a Re-ID network while freezing the underlying detection network.

This pipeline could expediently integrate state-of-the-art detection technologies and person Re-ID technologies that greatly simplifies the design difficulties of joint person search solutions.

3.5 Conclusions

Person search incorporating object detection and person re-identification is a very challenging task due to the difficulties of designing the detection network. This chapter proposed a deep-frozen transfer learning framework that trains a re-identification network underlying a pre-trained detector while keeping the entire detection network frozen. The re-identification network consisting of ATLnet and MDnet was proposed to learn multi-scale, multi-part appearance feature representation. The proposed framework implemented on multiple mainstream detectors, including YOLOv3, YOLOv4, Mask RCNN, and CenterNet has achieved remarkable performance improvements on two person search datasets CUHK-SYSU and PRW. The experimental results demonstrate that the listed pre-trained detection networks keep sufficient discriminative information for re-identification and the proposed shallow-designed ATLnet and MDnet could effectively extract discriminative features out from the frozen detection network. Our proposed framework, which is easy to integrate the state-of-the-art detection networks could greatly simplify the design of the joint person search solutions. The source code of our work (including training, evaluation, visualization, and a simple person search application) is available on github. [1]

1. https://github.com/RuaHU/TLfPS

CHAPTER 4

Online-Learning-Scene-Adaptive Data Association

4.1 Introduction

MOT aims to locate multiple objects of interest in a video sequence to create and maintain the continuous Spatio-temporal trajectories. Mainstream MOT algorithms adopt the tracking-by-detection paradigm; under this paradigm, MOT is modeled as a mid-level hybrid computer vision task that consists of and highly relies on other fundamental computer vision tasks, including object detection, appearance feature representation, and data association. Object detection and appearance feature representation (person re-identification) have their research communities. They have achieved remarkable improvement in recent years due to the more in-depth exploration of deep convolutional neural networks' powerful feature representation capabilities. However, data association, this essential problem is distinctly different from the other two tasks. It concerns processing multiple frame data; as a trade-off of the inference speed, most trivial association cues are dropped in the process of detection and appearance feature representation. As a result, data association does not have a well-defined learning-based solution so far. Traditional data association algorithms dominating the tracking fields yet lacking robustness in complicated scenarios.

We model data association as an inter-frame association problem and mainly consider two types of constraints:

— **Spatial-scale constraints**, the detection bounding boxes providing spatial-scale

constraints, like the bounding box distance and the bounding box shape, and size dissimilarity.

— **Feature space constraints**, the re-identification features providing appearance similarity measurement in feature space.

In Chapter 3, we proposed an end-to-end framework that incorporates object detection and re-identification. It provides a basic template for designing the multiple object tracking system. **This chapter in-depth analyzes the application limitations and advantages of re-identification features and proposes a novel approach to improve the association robustness.**

In many state-of-the-art MOT algorithms [68, 62, 63], re-identification features are deemed supplementary means to improve the robustness of the tracking system and hardly ever be used independently of spatial-scale cues to make association decisions. MAT[57] argue that, due to noisy partial-detections, and lack of temporal-spatial constraints, the re-identification feature is unreliable, time-consuming, and cannot address the false negatives for occluded and blurred objects. The reasons causing this underestimation are:

— The re-identification features in many tracking algorithms are not well adapted to cross-camera systems. When encountering large posture or appearance variation, re-identification features are prone to lose their function. This causes the unreliable.

— Well-trained re-identification features are designed to fit various scenes and appearance variations; they technically contain redundancies and lack significance for a specific tracking scene. The similarity distinction between real matching pairs and false matching pairs is insignificant that unreliable to make association decisions.

Our proposed person search network is designed for cross-camera system. It could tackle the first problem. This chapter is committed to discussing and solving the second problem— **how to employ re-identification features independently of spatial-scale constraints to make association decisions.**

The contents of this chapter are outlined as follows:

— In section 4.2, we formulate the data association problem.

— In section 4.3, we evaluate the spatial-scale and feature space association accuracies to illustrate the advantages and limitations of re-identification features.

— In section 4.4, we make a brief definition for the feature space and analyze the source of limitations. We propose a solution, named scene-adaptive data association, to improve the robustness of re-identification features.

— In section 4.5, we conduct experiments to validate the proposed association algorithm.

— In section 4.6, we conclude the advantages of our proposed solution.

4.2 Formulate the Data Association Problem

4.2.1 Formulate Object Detection

Given an image I_t at frame t, pedestrian detection is formulated by:

$$\mathbf{D} : I_t \mapsto D_t = \{det^1, det^2, ..., det^n\}_t \tag{4.1}$$

where $\mathbf{D}$ is the detection network, it maps the input image I_t to a set of detection D_t. $det^i = \{x, y, w, h\}_t^i$ is the i^{th} detection that composed of the center position $\{x, y\}_t^i$, the width and height $\{w, h\}_t^i$.

4.2.2 Formulate Re-Identification Feature Extraction

Given an image I_t, the re-identification feature representation of its detection D_t is formulated by:

$$\mathbf{R} : I_t, D_t \mapsto F_t = \{f_{det^1}, f_{det^2}, ..., f_{det^n}\}_t \tag{4.2}$$

where $\mathbf{R}$ is a re-identification network, it extracts appearance features from the image I_t for each bounding box of the detection D_t. f_{det^i} is the re-identification feature of the detection det^i.

4.2.3 Formulate Track List

The internal track list of the tracking system is formulated as:

$$T_{t-1} = \{track^1, track^2, ..., track^m\}_{t-1} \tag{4.3}$$

where $track_{t-1}^j = \{node^1, node^2, ...\}_{t-1}^j$ is a history trajectory, which composed of a set of nodes. $node^i = \{det, f_{det}\}^i$ consists of a detection det and its feature vector f_{det}. In the later contents, we use symbols $b_{track_{t-1}^j}$ and $f_{track_{t-1}^j}$ refer to the bounding box and appearance feature of $track_{t-1}^j$.

4.2.4 Formulate Motion Model

We adopt Kalman filter[6] to model the target's motion as a constant velocity model.

The motion state is defined in an 8-dimensional state space:

$$s_t = (x, y, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h}) \tag{4.4}$$

where (x, y) is the position, (γ, h) is the shape, here $\gamma = w/h$ is the width-height aspect ratio. $\{\dot{x}, \dot{y}, \dot{\gamma}, \dot{h}\}$ are the corresponding velocities. The Kalman's state transition equation is:

$$s_{t|t+1} = F_{t|t+1} s_t + u_{t+1} \tag{4.5}$$

where $F_{t|t+1}$ is a 8×8 state transition matrix. u_t is a 8-dimentional state noise vector. The observation equation is:

$$b_{t|t+1} = H s_{t|t+1} + v_{t+1} \tag{4.6}$$

where H is a 4×8 observation matrix. $b_{t|t+1}$ is a 4-dimensional predicted bounding box. v_t is a 4-dimentional observation noise vector.

The system noise is modeled by a 8×8 noise covariance matrix P_t. When starting a new track, the noise covariance matrix is initialized as a diagonal matrix:

$$P_0 = diag([\alpha h^2, \alpha h^2, \dot{\alpha}, \alpha h^2, \beta h^2, \beta h^2, \dot{\beta}, \beta h^2]) \tag{4.7}$$

where h is the height of bounding box. The parameters α, β, $\dot{\alpha}$, $\dot{\beta}$ are hyper-parameters that based on specific scenes.

More details of the Kalman filter implementation and parameter selection could refer to [68, 69] and their projects.

4.2.5 Formulate Data Association

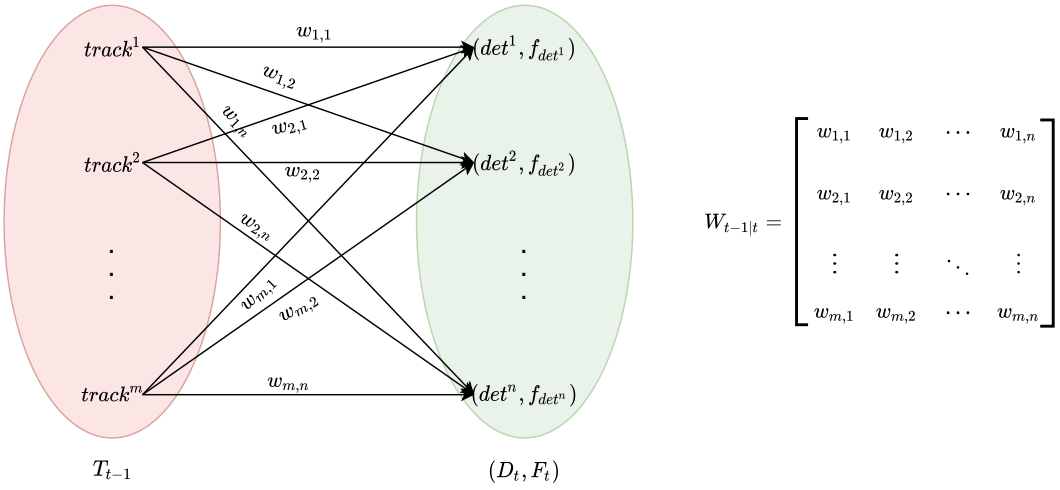


Figure 4.1 Fully connected bipartite graph.

See Fig. 4.1; The association problem could be modeled as a weighted bipartite graph solving problem:

$$G_{t-1|t} = ([T_{t-1}, (D_t, F_t)]; E_{t-1|t}; W_{t-1|t}) \tag{4.8}$$

where $G_{t-1|t}$ is the weighted bipartite graph between frame $t-1$ and frame t. The track list T_{t-1} and the detection (D_t, F_t) are the vertex of the graph. $E_{t-1|t}$ is a collection of edges connecting tracks T_{t-1} and detection D_t. $W_{t-1|t}$ is the corresponding weights of $E_{t-1|t}$. In an association problem, the graph is fully connected and has $m \times n$ edges. The corresponding weights (weight of edges) $W_{t-1|t}$ are denoted as a matrix: $W_{t-1|t} \in \mathbb{R}^{m \times n}$.

Solving the weighted bipartite graph is to search the maximum pairs of matches that maximize the sum of weights. The solver is the Hungarian Algorithm (HA) [98]:

$$\mathbf{HA} : W_{t-1|t} \mapsto M_{t-1|t} \tag{4.9}$$

where $M_{t-1|t}$ is the matches between T_{t-1} and D_t. There are $min(m, n)$ matches in maximum and could be denoted as:

$$M_{t-1|t} = \{match^l_{t-1|t}\}_{l=1}^{min(m,n)} \tag{4.10}$$

where $match^l_{t-1|t} = (i, j)$ is the l^{th} matching edge of the graph $G_{t-1|t}$ that connects the i^{th} track and the j^{th} detection. Usually, we will set a threshold to filter out low-weight edges. The overall matches will equal to or less than $min(m, n)$. It worth noting that, according to the definition of **HA**, when mis-detection happening, tracks are prone to switch IDs.

4.3 Multiple Cues Association

As discussed, solving the data association problem is the building and solving of the weight matrix. This section will define the weight matrix $W_{t-1|t} \in \mathbb{R}^{m \times n}$ using spatial-scale constraints and feature space constraints and evaluate their association accuracies.

4.3.1 The Expression of Weight Matrix

Re-Identification Features Similarity Matrix

The appearance similarity matrix between tracks T_{t-1} and detection F_t could be written as:

$$W^{reid}_{t-1|t} = \{w^{reid}_{i,j}\}_{i=1,j=1}^{m,n} \tag{4.11}$$

where:

$$w^{reid}_{i,j} = \frac{\sum f_{track^i_{t-1}} \dot{f}_{det^j}}{\left\| f_{track^i_{t-1}} \right\|_2 \| f_{det^j} \|_2} \tag{4.12}$$

is the cosine similarity between the track's re-identification feature $f_{track^i_{t-1}}$ and the detection re-identification feature f_{det^j}. $\|\cdot\|_2$ is $l2$ norm.

Intersection over Union

The intersection over union (IoU) between tracks T_{t-1} and detection D_t is:

$$W_{t-1|t}^{IoU} = \{w_{i,j}^{IoU}\}_{i=1,j=1}^{m,n} \tag{4.13}$$

where:

$$w_{i,j}^{IoU} = \frac{a_i \bigcap a_j}{a_i + a_j - a_i \bigcap a_j} \tag{4.14}$$

is the IoU between the track's prediction bounding box of $b_{track_{t-1|t}^i}$ and detection det_t^j. a_i is the area of bounding box $b_{track_{t-1|t}^i}$. a_j is the area of bounding box det_t^j. $a_i \bigcap a_j$ is the intersection area. The range of IoU is in $[0, 1]$, once the two bounding boxes distant from each other, the intersection area becomes 0, and the IoU becomes 0. This limit the application of IoU in fast-moving scene.

Mahalanobis Distance

DeepSORT [62] propose to use Mahalanobis distance instead of IoU to model the distance between bounding boxes:

$$W_{t-1|t}^{dist} = \{w_{i,j}^{dist}\}_{i=1,j=1}^{m,n} \tag{4.15}$$

where:

$$\begin{aligned} w_{i,j}^{dist} &= 1 - \frac{2}{1 + \exp(-d_{i,j})} \\ d_{i,j} &= (det_t^j - b_{track_{t-1|t}^i})^T (H^T P_{t-1|t} H)^{-1} (det_t^j - b_{track_{t-1|t}^i}) \end{aligned} \tag{4.16}$$

is the Mahalanobis distance between the track's predicted bounding box $b_{track_{t-1|t}^i}$ and the detection det_t^j. $P_{t-1|t}$ is the predicted noise covariance matrix. It worth noting that we limit the distance weight between $[-1, 1]$.

Join Re-identification Features Similarity and Mahalanobis Distance

Most state-of-the-art trackers [68, 69] using join detection and re-identification (embedding) paradigm to model the weight matrix. The solution fuses the appearance similarity matrix and distance matrix by:

$$W_{t-1|t}^{jde} = \alpha W_{t-1|t}^{reid} + (1 - \alpha) W_{t-1|t}^{dist} \tag{4.17}$$

where $\alpha \in [0, 1]$ is used to adjust the importance of different constraints. When $\alpha = 0$, the weight matrix equals the distance matrix $W_{t-1|t}^{dist}$. When $\alpha = 1$, the weight matrix equals the similarity matrix.

4.3.2 The Association Algorithm

It is pretty chaotic that we listed many equations to formulate the association problem. To clarify the association problem, we summarize the whole process in the following algorithm.

Algorithm 1: Association()

Input : D_t, F_t, T_{t-1}

Output: $M_{t-1|t}$

`/* get appearance similarity matrix and distance matrix */`

foreach $track^i_{t-1} \in T_{t-1}$, $det^j_t \in D_t$ **do**

$W^{reid}_{t-1|t} = \{w^{reid}_{i,j}\}^{n,m}_{i=1,j=1}$;

$w^{reid}_{i,j} = \frac{\sum f_{track^i_{t-1}} f_{det^j_t}}{\left\| f_{track^i_{t-1}} \right\|_2 \left\| f_{det^j_t} \right\|_2}$;

$W^{dist}_{t-1|t} = \{w^{dist}_{i,j}\}^{n,m}_{i=1,j=1}$;

$w^{dist}_{i,j} = 1 - \frac{2}{1+\exp(-d_{i,j})}$;

KalmanFilter : $track^i_{t-1} \mapsto track^i_{t-1|t}$;

$d_{i,j} = (det^j_t - b_{track^i_{t-1|t}})^T (H^T P_{t-1|t} H)^{-1} (det^j_t - b_{track^i_{t-1|t}})$;

end

`/* fuse multiple cues */`

$W_{t-1|t} = \alpha W^{reid}_{t-1|t} + (1-\alpha) W^{dist}_{t-1|t}$;

`/* solve the weight matrix by using HA */`

HA : $W_{t-1|t} \mapsto M_{t-1|t}$

This algorithm composes two steps: **i**. calculate the weight matrix $W_{t-1|t}$. **ii**. solve the matrix by using the Hungarian Algorithm (HA). The output of this algorithm are matches that pairwise connecting the tracks and detections.

4.3.3 Verify the Robustness of the Re-Identification Feature Association

Intuitively, re-identification features having no spatial constraints could adapt to more challenging tracking scenes, i.e., tracking the fast-moving targets or re-tracking lost targets. In this section, we will set experiments to prove the robustness of using re-identification features for association.

We use the chapter 3 proposed person search network—CenterNet-MDnet as the detection and re-identification network, and the algorithm 1 as association algorithm. We select

Table 4.1 MOT17 Training Sequences

MOT17 Training Sequences							
Name	FPS	Resolution	Frames	Tracks	Boxes	Viewpoint	Camera
MOT17-02	30	1920x1080	600	49	17833	medium	static
MOT17-04	30	1920x1080	1050	80	47557	high	static
MOT17-05	14	640x480	837	124	6818	medium	moving
MOT17-09	30	1920x1080	525	25	5257	low	static
MOT17-10	30	1920x1080	654	54	12318	medium	moving
MOT17-11	30	1920x1080	900	67	9174	medium	moving
MOT17-13	25	1920x1080	750	68	11450	high	moving
Total			5316	512	110407		

different parameters to evaluate the association accuracies of different association cues.

— The parameter α in algorithm 1 is set to 1 to evaluate the accuracy of purely using feature space constraints.
— The parameter α in algorithm 1 is set to 0 to evaluate the accuracy of purely using the spatial-scale constraints.
— The parameter α in algorithm 1 is set to 0.5 to evaluate the accuracy of using join spatial-scale constraints and feature space constraints.

The association accuracy is defined as:

$$\begin{aligned} acc &= \frac{M_{t-1|t}^{true}}{M_{t-1|t}^{true} + M_{t-1|t}^{false}} \\ M_{t-1|t}^{true} &= M_{t-1|t} \cap M_{t-1|t}^{gt} \\ M_{t-1|t}^{false} &= (M_{t-1|t} \cup M_{t-1|t}^{gt}) - M_{t-1|t}^{gt} \end{aligned} \tag{4.18}$$

where $M_{t-1|t}$ is the output of the Hungarian Algorithm (the output of algorithm 1). Note that we set a $thresh = 0.5$ for the weight matrix $W_{t-1|t}$, the elements less than 0.5 will not be taken into consideration. $M_{t-1|t}^{gt}$ is the ground truth matches given by datasets. $M_{t-1|t}^{true}$ are true positive matches. $M_{t-1|t}^{false}$ are false positive matches (should be matched but not matched) and false negative (shouldn't be matched but matched) matches.

There is no specific dataset for evaluating the association accuracy. We use the labeled training sequences of MOT17 for the evaluation. The detailed information of the training sequences of MOT17 is shown in Table. 4.1. The association accuracies of using different association constraints and different frame interval are shown in Fig. 4.2. There are 7 MOT17 sequences that get evaluated. Among them, the videos MOT17-02, MOT17-04, MOT17-09 are captured by static cameras. MOT17-05, MOT17-10, MOT17-11, MOT17-

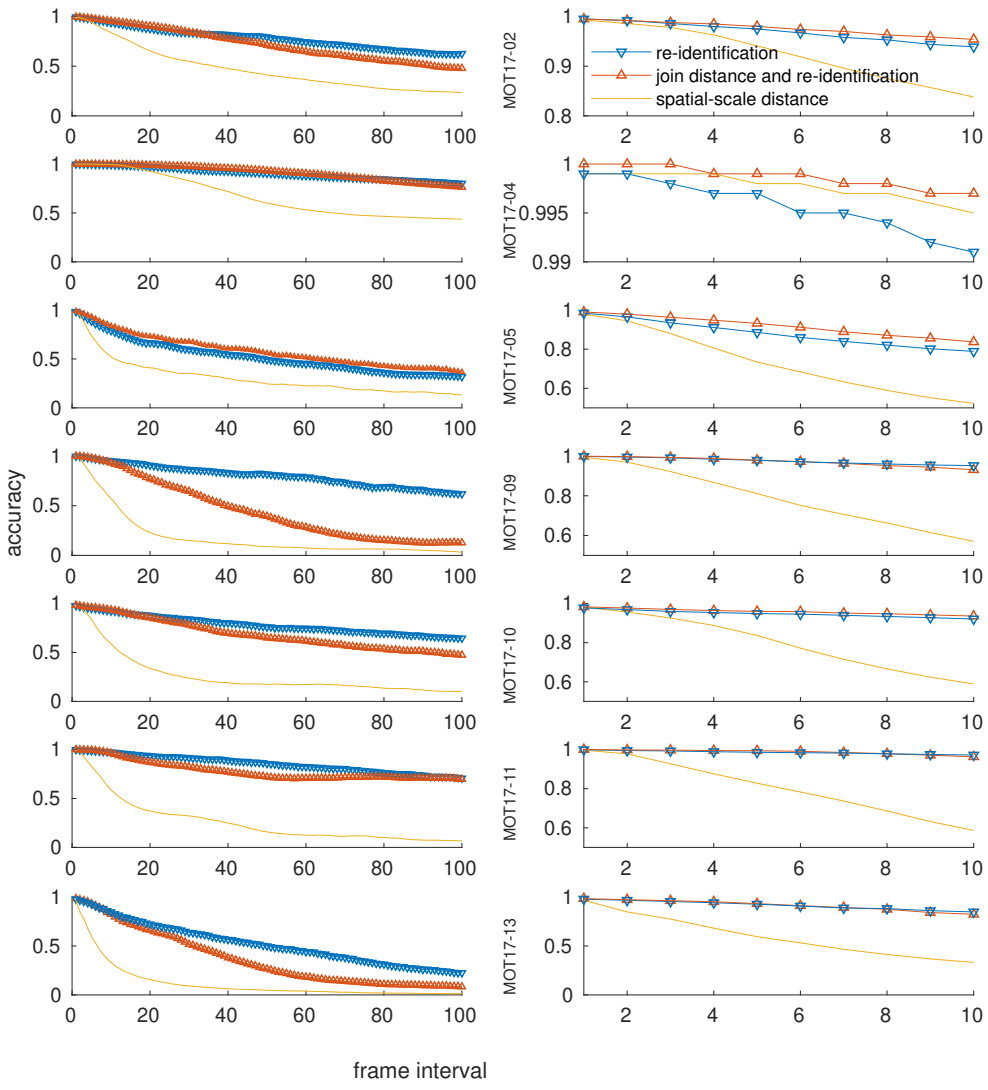


Figure 4.2 Association accuracy by using different weight matrix and different frame interval.

13 are captured by mobile cameras.

It could be seen that when the frame interval is small, i.e., 1, the association accuracy is high, it could provide reliable association accuracy for the inter-frame association problem. Especially, the re-identification feature association accuracy exceeds 97% on all video sequences. The accuracy of join distance and re-identification exceeds 98% on all videos. When the frame interval increases, re-identification feature association shows its advantages in robust association for low frame rate videos. Intuitively and experimentally, applying spatial-scale constraints for association requires a strong assumption that the object's inter-frame motion is small. Once frame interval increases or the camera changes its position or direction, the association accuracy degrades quickly; thus, lacking robustness. Feature space association is more robust across all frame rates and all scenes.

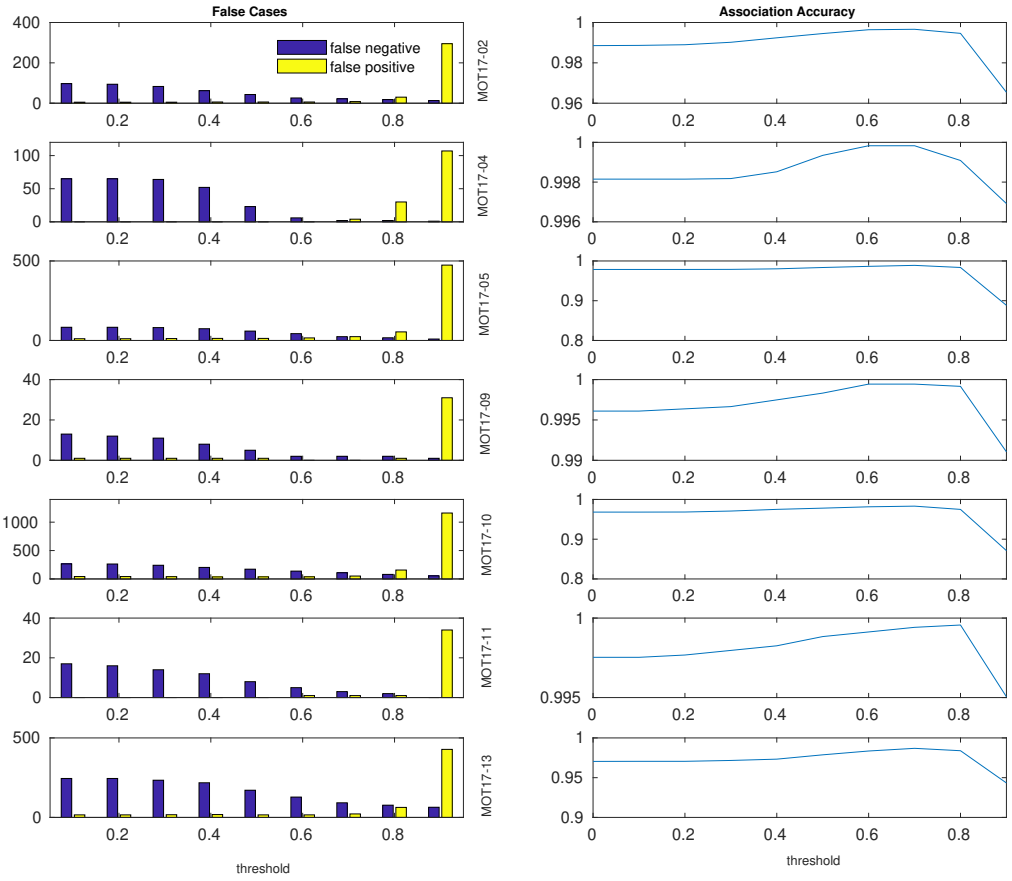


Figure 4.3 Re-identification feature Association accuracy with different threshold.

We conduct the second group of experiments to study the effect of selecting various thresholds to the re-identification feature association accuracy. We fixed the above experiment

threshold to 0.5. Here, we change the threshold from 0 → 1 with 0.1 interval and fix the frame interval to 1. The results are shown in Fig. 7.6. We plot the number of false-positive and false-negative cases. It could be seen from the figure that when the threshold increases, the number of false-positive cases increases, and the false-negative decreases. When the threshold is 0.7, the total number of false match cases are minimal. In this case, the inter-frame association accuracy of using re-identification features exceeds 98%.

Our experiments prove that re-identification features can provide high association accuracy. It has an incomparable advantage comparing spatial-scale bounding box association in low frame rate videos. However, the existing problem is that when a target lost its detection in the next frame. The re-identification feature cannot give sufficient discrimination to judge whether the target is lost or not. In the left side of Fig. 4.4, given an association threshold 0.7, when the query target lost its detection in $frame2$, it will match a wrong target, which has similarity 0.81 that greater than the threshold 0.7. The two targets are well distant in the spatial-scale but failed in feature space. A robust tracking system needs to frequently tackle these lost tracks, which beyond the capability of re-identification features. It is inappropriate to use the re-identification feature independently of the spatial-scale to make association decisions.

Consider that if the re-identification feature provides discriminative similarity, as shown in the right side of Fig. 4.4. When the query target lost its detection in the next frame, the similarity to the negative match is 0.5, which less than 0.7. In this case, the re-identification feature making correct association decisions could adapt to more complicated scenes.

Improving the discriminativeness of the re-identification feature and solving the association problem in the feature-space is a problem to be solved in this chapter.

FALSE: the query person is associated with an incorrect target

TRUE: the query person is not associated with any target

Figure 4.4 Illustrate the limitation of the re-identification feature.

4.4 Proposed Scene-Adaptive Data Association

As afore discussed, the re-identification feature has incomparable advantages comparing spatial-scale bounding box association in low frame rate videos. It is widely used in many state-of-the-art tracking algorithms. Still, it plays a trivial role in the association problem of fast-moving scenes because it lacks discrimination to make association decisions. This section proposes a solution to tackle this problem.

4.4.1 Re-identification-feature space and Association-feature space

Cross camera person re-identification and intra-camera, inter-frame data association are two distinct tasks. Person re-identification focus on the profile of pedestrians' appearance, i.e., the color, style, detailed textures of the pedestrian's hair, clothes, shoes, trousers, etc. Pedestrians' cross-camera posture changing greatly is less important. Whereas in the intra-camera, inter-frame data association problem, pedestrians posture is more essential than appearance to distinguish different identities. Directly utilizing re-identification features for data association is inappropriate. We argue that the feature space in different camera systems is different. It is impossible to train a general feature space that could adapt well to all scenarios. It is necessary to map the cross camera 'general-purpose' re-identification feature from re-identification-feature space to 'specific-purpose' association-feature space.

In order to better discuss our point of view, we give a brief description of the re-identification-feature space. We denote a re-identification-feature space as: $\boldsymbol{\chi}^{reid} = \{base_i\}_{i=1}^{n}$. Where $base_i$ is the i^{th} *basevector* of the feature space, it is an abstract appearance descriptor, i.e.,

$$base_i \in \{hair^{black}, hair^{white}, hair^{long}, hair^{short}, coat^{yello}, coat^{green}, mask^{mouth}...\} \quad (4.19)$$

where $hair^{black}$ means the pedestrian's *black hair*. A well trained re-identification-feature space will contain as many distinct types of abstract appearance descriptors as possible. The higher rank of the feature space, the better representation capability of the feature space. A re-identification feature could be defined as $f^{reid} = \{\gamma_1, \gamma_2, ..., \gamma_n\}$. Where γ_i is the intensity of $base_i$. For example, given three pedestrians p_1, p_2, p_3, and their corresponding re-identification features:

- $f_{p_1}^{reid} = \{0.9, 0.0, 0.0, 0.9, 0.0, 0.8, 0.0, ...\}$ means *a person with short (*0.9 *intensity) black (*0.9 *intensity) hair, and wears green (*0.8 *intensity) coat.*
- $f_{p_2}^{reid} = \{0.7, 0.0, 0.0, 0.8, 0.0, 0.9, 0.0, ...\}$ means *a person with short (*0.7 *intensity) black (*0.8 *intensity) hair, and wears green (*0.8 *intensity) coat.*

— $f_{p_3}^{reid} = \{0.0, 0.8, 0.8, 0.0, 0.9, 0.0, 0.0, ...\}$ means *a person with long (*0.8 *intensity) white (*0.8 *intensity) hair, and wears yellow (*0.9 *intensity) coat.*

In the re-identification feature space, pedestrians p_1 and p_2 have a close appearance; they are less distant in feature space. Person p_1 and p_3 distinguishing in appearance will distant in feature space. It is easy to exclude the association between p_3 and p_1 but isn't easy to decide whether p_1 and p_2 are from the same identity because may multiple pedestrians own similar hairstyles and clothes in a scene. It lacks discriminative to make associated decisions. In Fig. 4.4, the negative candidate person has close dresses to the query person. It is impossible to distinguish the two pedestrians on appearance. The more significant differences are the wearing *mask* and the varying posture. In an association problem, we intend to minimize the importance of the common feature *bases* and increase the importance of the decisive feature *bases*. For example, given two person p_1, p_2 with association feature:

— $f_{p_1}^{reid} = \{0.9, 0.0, 0.0, 0.9, 0.0, 0.8, 4.0, ...\}$ means *a person with short (*0.9 *intensity) black (*0.9 *intensity) hair, wears green (*0.8 *intensity) coat and a mouth mask (*4.0 *intensity).*

— $f_{p_2}^{reid} = \{0.7, 0.0, 0.0, 0.8, 0.0, 0.9, 0.0, ...\}$ means *a person with short (*0.7 *intensity) black (*0.8 *intensity) hair, wears green (*0.9 *intensity) coat.*

For $f_{p_1}^{reid}$, the intensity of $mask^{mouth}$ dominates the feature vector; the two pedestrians could well distinct by this descriptor. A well-trained re-identification feature space will have a large number of redundant feature *bases*, the feature *base*—$mask^{mouth}$ exists but contributes low intensity; thus, it can not provide sufficient discriminative to distinguish the two pedestrians. To deal with this problem, we define a converter to convert a re-identification feature vector from a re-identification feature space to an associated feature space. The principle is to re-evaluate the importance of each feature *base*, increases the importance of decisive feature *bases*, and decrease the importance of other redundant common feature *bases*.

We define a converter that converts the re-identification feature vectors from re-identification-feature space to association-feature space:

$$\boldsymbol{\chi}^{asso} = \mathbf{W}^{reid|asso}\boldsymbol{\chi}^{reid} + \mathbf{B} \tag{4.20}$$

Where $\mathbf{W}^{reid|asso} \in \mathbb{R}^{n,m}$ is the conversion matrix. $\boldsymbol{\chi}^{asso}$ is an association feature space that has m *bases*. $\mathbf{B}$ is a m-dimensional bias vector.

We intend to find out the conversion matrix $\mathbf{W}^{reid|asso}$ and the bias vector $\mathbf{B}$. Here, $\mathbf{T} = \{\mathbf{W}^{reid|asso}, \mathbf{B}\}$ is called converter. In principle, pre-training a converter for each specific scene generally requires a large amount of labeled data, which is inefficient and impossible. In this section, we propose an online-learning method to adaptively update the weights of $\mathbf{T} = \{\mathbf{W}^{reid|asso}, \mathbf{B}\}$.

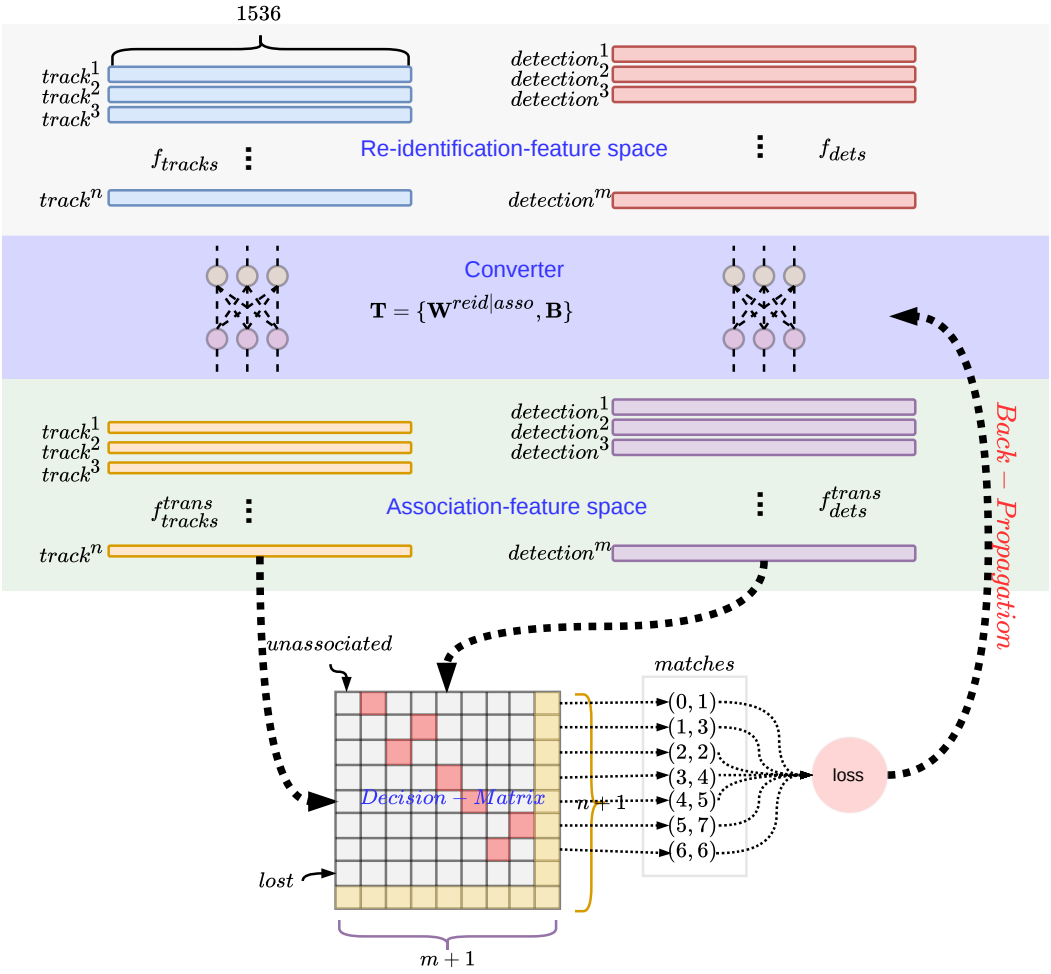


Figure 4.5 The Architecture of Scene-Adaptive Data Association.

4.4.2 Formulate the Architecture of Scene-Adaptive Data Association

Suppose there are n tracks $T_{t-1} = \{track^1, track^2, ..., track^n\}_{t-1}$ and m detection $D_t = \{det^1, det^2, ..., det^m\}$. We first define the association matrix (weight matrix by using converted features) as:

$$\begin{aligned} W_{t-1|t} &= \{w_{i,j}\}_{i=1,j=1}^{n,m} \\ w_{i,j} &= \frac{\sum f_{track^i}^{trans} f_{det^j}^{trans}}{\left\| f_{track^i}^{trans} \right\|_2 \left\| f_{det^j}^{trans} \right\|_2} \end{aligned} \tag{4.21}$$

where:

$$\begin{aligned} f_{track^i}^{trans} &= \mathbf{T}_{t-1|t}(f_{track^i}) = f_{track^i} \mathbf{W}_{t-1|t}^{reid|asso} + \mathbf{B}_{t-1|t} \\ f_{det^j}^{trans} &= \mathbf{T}_{t-1|t}(f_{det^j}) = f_{det^j} \mathbf{W}_{t-1|t}^{reid|asso} + \mathbf{B}_{t-1|t} \end{aligned} \tag{4.22}$$

where f_{track^i} is the i^{th} track's re-identification feature. f_{det}^{j} is the j^{th} detection's re-identification feature. f^{trans} is the converted feature vector.

We intend to make association decisions by solving the weight matrix $W_{t-1|t}$. For the purpose of better express the lost tracks and unassociated objects, we use the same idea as [99] that append an extra column and row of scalar values to expand the weight matrix:

$$\bar{W}_{t-1|t} = \begin{pmatrix} & & \alpha \\ & W_{t-1|t} & \vdots \\ & & \alpha \\ \alpha & \dots & \alpha \quad \alpha \end{pmatrix} \tag{4.23}$$

where α is a scalar. We perform softmax for each row and each column of the matrix $\bar{W}_{t-1|t}$. The softmax matrices are denoted as:

$$\begin{aligned} \tilde{W}_{t-1|t}^{row} &= softmax^{row}(\bar{W}_{t-1|t}\tau) \\ \tilde{W}_{t-1|t}^{col} &= softmax^{col}(\bar{W}_{t-1|t}\tau) \end{aligned} \tag{4.24}$$

where τ is a trainable parameter. The two matrices $\tilde{W}_{t-1|t}^{row}$ and $\tilde{W}_{t-1|t}^{col}$ are called *Decision-Matrices*. Supposing $w_{i,j}^{row}$ and $w_{i,j}^{col}$ are the i^{th} row and j^{th} column of $\tilde{W}_{t-1|t}^{row}$ and $\tilde{W}_{t-1|t}^{col}$, according to the value, we could make three association decisions:

- **match**: if $i < n, j < m$ and $w_{i,j}^{row} > thresh$, $w_{i,j}^{col} > thresh$. Then (i, j) is the index of a match that connect the i^{th} track and the j^{th} detection, the corresponding match weight is: $(w_{i,j}^{row}, w_{i,j}^{col})$.
- **unassociated track**: if $j = m$ and $w_{i,j}^{row} > thresh$, then the i^{th} track marked as an unassociated track.
- **unassociated detection**: if $i = n$ and $w_{i,j}^{col} > thresh$, then the j^{th} detection marked as an unassociated detection.

4.4.3 Loss and Back-Propagation

The objective of the loss function is to maximize the weight of matched points and minimize the weight of unmatched points of the two $Decision - Matrices$:

$$w_{i,j} = \begin{cases} 1 & if\ w_{i,j}\ matched \\ 0 & otherwise \end{cases} \tag{4.25}$$

Supposing there are $N\ mathes = \{(w_i^{row}, w_i^{col})\}_{i=1}^{N}$. The loss function is:

$$loss = -\frac{1}{2N}\sum_{i=1}^{N}(\log w_i^{row} + \log w_i^{col}) \tag{4.26}$$

The whole process is differentiable, the Back-Propagation equations are written as:

$$\begin{aligned}
\mathbf{W}_{t|t+1}^{reid|asso} &= \mathbf{W}_{t-1|t}^{reid|asso} + \eta \frac{\delta_{loss}}{\delta_{\mathbf{W}_{t-1|t}^{reid|asso}}} \\
\mathbf{B}_{t|t+1} &= \mathbf{B}_{t-1|t} + \eta \frac{\delta_{loss}}{\delta_{\mathbf{B}_{t-1|t}}} \\
\alpha &= \alpha + \eta \frac{\delta_{loss}}{\delta_{\alpha}} \\
\tau &= \tau + \eta \frac{\delta_{loss}}{\delta_{\tau}}
\end{aligned} \tag{4.27}$$

where η is the learning rate.

As discussed in section 4.3.3, inter-frame association in both the spatial-scale and the feature space has very high accuracy. The data fed back to the converter is solid and safe. In principle, this self-iterative online-training approach is also solid.

4.4.4 Algorithm

Again, we listed many equations, and it isn't easy to understand the logic of all of them. In order to clarify all the aforementioned formulas, in this subsection, we give the complete association algorithm.

Initialization

When initializing training/tracking, the conversion matrix $\mathbf{W}_{0|1}^{reid|asso} \in \mathbb{R}^{n \times n}$ initialized as an identity matrix. Where n equals to the dimension of re-identification feature vector. The bias vector $\mathbf{B}_{0|1} \in \mathbb{R}^n$ is a zero vector.

Algorithm 2: Initialization()

Output: $T_{t-1}, \mathbf{D}, \mathbf{R}, \mathbf{W}_{0|1}^{reid|asso}, \mathbf{B}_{0|1}$

Initialize Tracks $T_{t-1} = \{\}$;

Initialize the detection network $\mathbf{D}$;

Initialize the re-identification network $\mathbf{R}$;

Initialize the Converter: $\mathbf{W}_{0|1}^{reid|asso} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}; \mathbf{B}_{0|1} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

Generating Decision-Matrices

This algorithm forwards the computation of converter and outputs two $Decision-Matrices$.

Algorithm 3: DecisionMatric()

Input : I_t, $\mathbf{D}$, $\mathbf{R}$, $\mathbf{W}_{t-1|t}^{reid|asso}$, $\mathbf{B}_{t-1|t}$

Output: $\tilde{W}_{t-1|t}^{row}$, $\tilde{W}_{t-1|t}^{col}$, D_t, F_t

$\mathbf{D} : I_t \mapsto D_t = \{det^1, det^2, ..., det^m\}_t$;

$\mathbf{R} : I_t, D_t \mapsto F_t = \{f_{det^1}, f_{det^2}, ..., f_{det^m}\}_t$;

`/* Compute association-features for tracks */`

$f_{track^i}^{trans} = \mathbf{T}(f_{track^i}) = f_{track^i}\mathbf{W}^{reid|asso} + \mathbf{B}$;

`/* Compute association-features for detection */`

$f_{det^j}^{trans} = \mathbf{T}(f_{det^j}) = f_{det^j}\mathbf{W}^{reid|asso} + \mathbf{B}$;

`/* Compute weight matrix */`

$W_{t-1|t} = \{w_{i,j}\}_{i=1,j=1}^{n,m}$;

where $w_{i,j} = \frac{\sum f_{track^i}^{trans} f_{det^j}^{trans}}{\left\|f_{track^i}^{trans}\right\|_2 \left\|f_{det^j}^{trans}\right\|_2}$;

`/* Compute expansion matrix */`

$$\bar{W}_{t-1|t} = \begin{pmatrix} & & \alpha \\ & W_{t-1|t} & \vdots \\ & & \alpha \\ \alpha & \dots & \alpha \quad \alpha \end{pmatrix};$$

`/* perform softmax in the row and col direction */`

$\tilde{W}_{t-1|t}^{row} = softmax^{row}(\bar{W}_{t-1|t}\tau)$;

$\tilde{W}_{t-1|t}^{col} = softmax^{col}(\bar{W}_{t-1|t}\tau)$;

Scene-Adaptive Data Association Algorithm

This algorithm consists of three steps:

1 Solving the $Decision - Matrices$ to fetch the matching pairs. Note that, at the very beginning of training, this step generates very few matches. Most of the matches will be generated by the second step.

2 Solving the association problem for unassociated tracks and detections by using the algorithm. 1.

3 Getting the loss and back propagate the gradients to the converter to update its weights.

Algorithm 4: Scene-Adaptive Data Association

Input: Video $= \{I_t\}_{t=1}^{n}$

T_0, $\mathbf{D}$, $\mathbf{R}$, $\mathbf{W}_{0|1}^{reid|asso}$, $\mathbf{B}_{0|1}$=`Initialization()`;

while *recived frame* I_t **do**

  $\tilde{W}_{t-1|t}^{row}$, $\tilde{W}_{t-1|t}^{col}$, D_t, F_t=`DecisionMatrix(`I_t, $\mathbf{D}$, $\mathbf{R}$, $\mathbf{W}_{t-1|t}^{reid|asso}$, $\mathbf{B}_{t-1|t}$`)`;

  $matches^{reid} = \{\}$;

  foreach $w_{i,j}^{row} \in \tilde{W}_{t-1|t}^{row}$, $w_{i,j}^{col} \in \tilde{W}_{t-1|t}^{col}$ **do**

    if $w_{i,j}^{row} > thresh$ **then**

      if $w_{i,j}^{col} > thresh$ **then**

        $(w_{i,j}^{row}, w_{i,j}^{col}) \rightarrow matches^{reid}$

      end

    end

  end

  `/* associate unassociated tracks and detection */`

  $matches^{join}$=`Association(`D_t^{unasso}, F_t^{unasso}, T_{t-1}^{unasso}`)`;

  $matches = matches^{reid} \cup matches^{join}$;

  `/* get loss */`

  $loss = -\frac{1}{2N}\sum_{i=1}^{N}(\log w_i^{row} + \log w_i^{col})$;

  `/* back propogation */`

  $\mathbf{W}_{t|t+1}^{reid|asso} = \mathbf{W}_{t-1|t}^{reid|asso} + \eta\frac{\delta_{loss}}{\delta_{\mathbf{W}_{t-1|t}^{reid|asso}}}$;

  $\mathbf{B}_{t|t+1} = \mathbf{B}_{t-1|t} + \eta\frac{\delta_{loss}}{\delta_{\mathbf{B}_{t-1|t}}}$;

  $\alpha = \alpha + \eta\frac{\delta_{loss}}{\delta_{\alpha}}$;

  $\tau = \tau + \eta\frac{\delta_{loss}}{\delta_{\tau}}$

end

4.5 Evaluation

In this section, we will conduct experiments to illustrate the effectiveness of the proposed Scene-Adaptive Converter. The configurations of the experiments are as follow: The detection network and the re-identification network are CenterNet-MDnet. The parameter α in equation 4.23 is set to 0.5. The initial value of parameter t in equation 4.24 is set to 10. The learning rate η is 0.2. The *thresh* in algorithm 4 is set to 0.9.

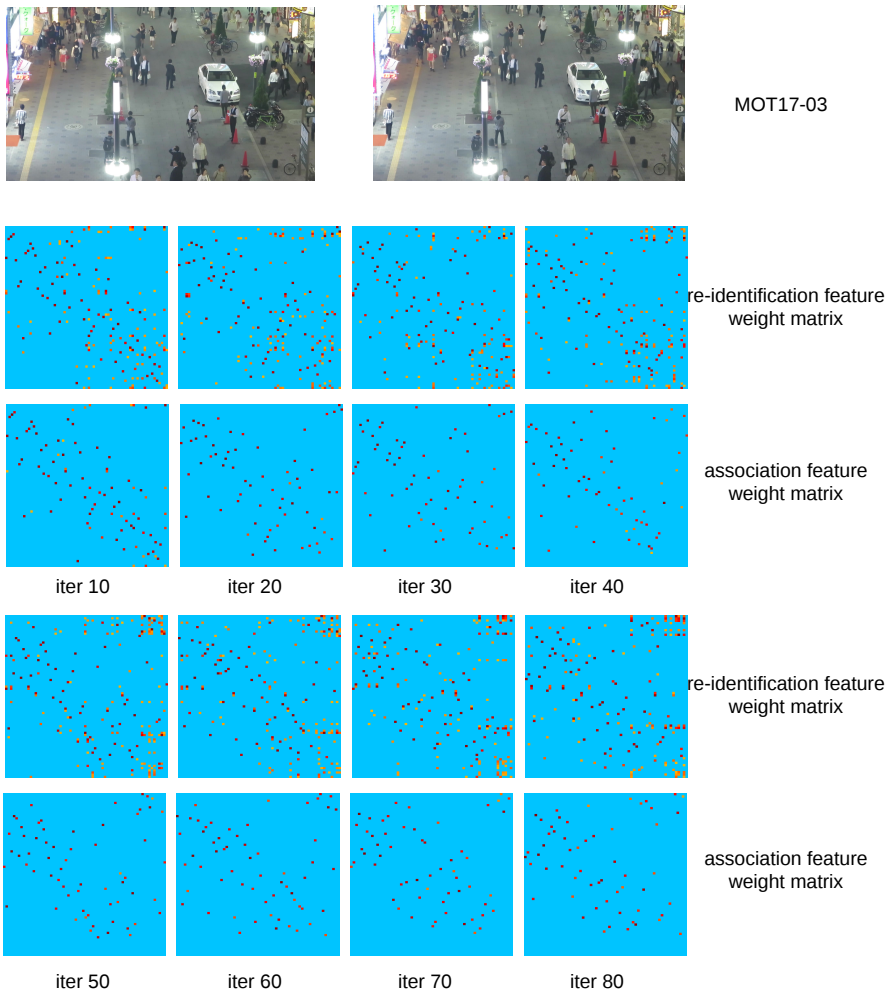


Figure 4.6 Illustrate the updating of the weight matrix.

4.5.1 Visualize The Updating of the Association Matrix

Fig. 4.6 shows the updating of the association matrix for crowded tracking scene MOT17-03. We set a threshold to make the association matrix background clean. The similarity lower than 0.7 was set to 0. It could be seen that after several iterations, the association feature weight matrix (AFWM) becomes clear (each row or each col of the matrix has at most 1 association point) than the re-identification feature weight matrix (RFWM). Based on the AFWM, we could make association decisions. For example, in iteration 80, the last few rows of the weight matrix are the similarity between lost tracks and detection. In the AFWM, the lost tracks are not associated with any detection. The tracks are marked as lost. In RFWM, the lost tracks have multiple association points with similarities great than 0.7, failed to make association decisions.

Fig. 4.7 shows the proportion of $matches^{reid}$ and $matches^{join}$. At the beginning of the

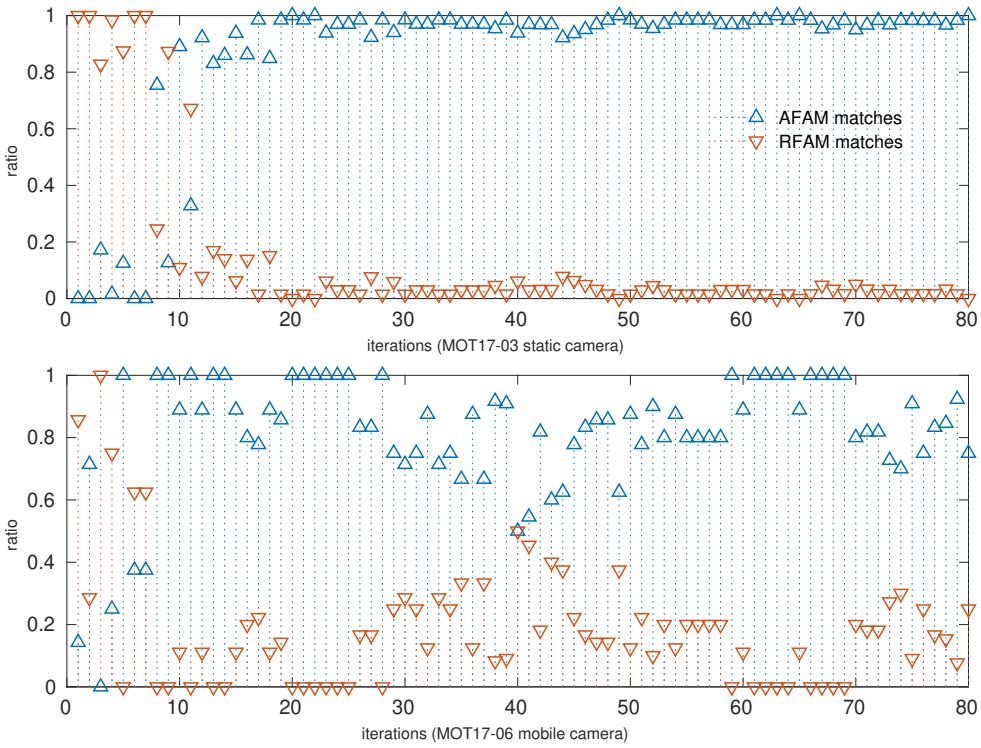


Figure 4.7 Proportion of matches.

tracking process, most matches get from bounding box association; re-identification features are lack significance. With the training/tracking process forwards, the association feature gets trained and shows its power to make association decisions. It worth noting that we get the $matches^{reid}$ from the two $Decision-Matrices$; the matches require significance. In this experiment, we set the threshold of the $Decision-Matrix$ to 0.9, which means the similarity between positive matches must significantly larger than all other negative matches. This is a stringent requirement to ensure the association's reliability, re-identification features lacking significance cannot satisfy this requirement.

4.5.2 Visualize Fast Moving Scene

When camera rotates fast, spatial-scale bounding box association will lose its function because it does not satisfy the spatial-scale constraint. This will result in loss tracking. Fig. 4.8 shows the association results using our proposed method; bounding boxes with the same color indicates the same person. The detailed association weight matrices are shown. It could be seen that the AFWM has much discriminative similarity than RFWM. The

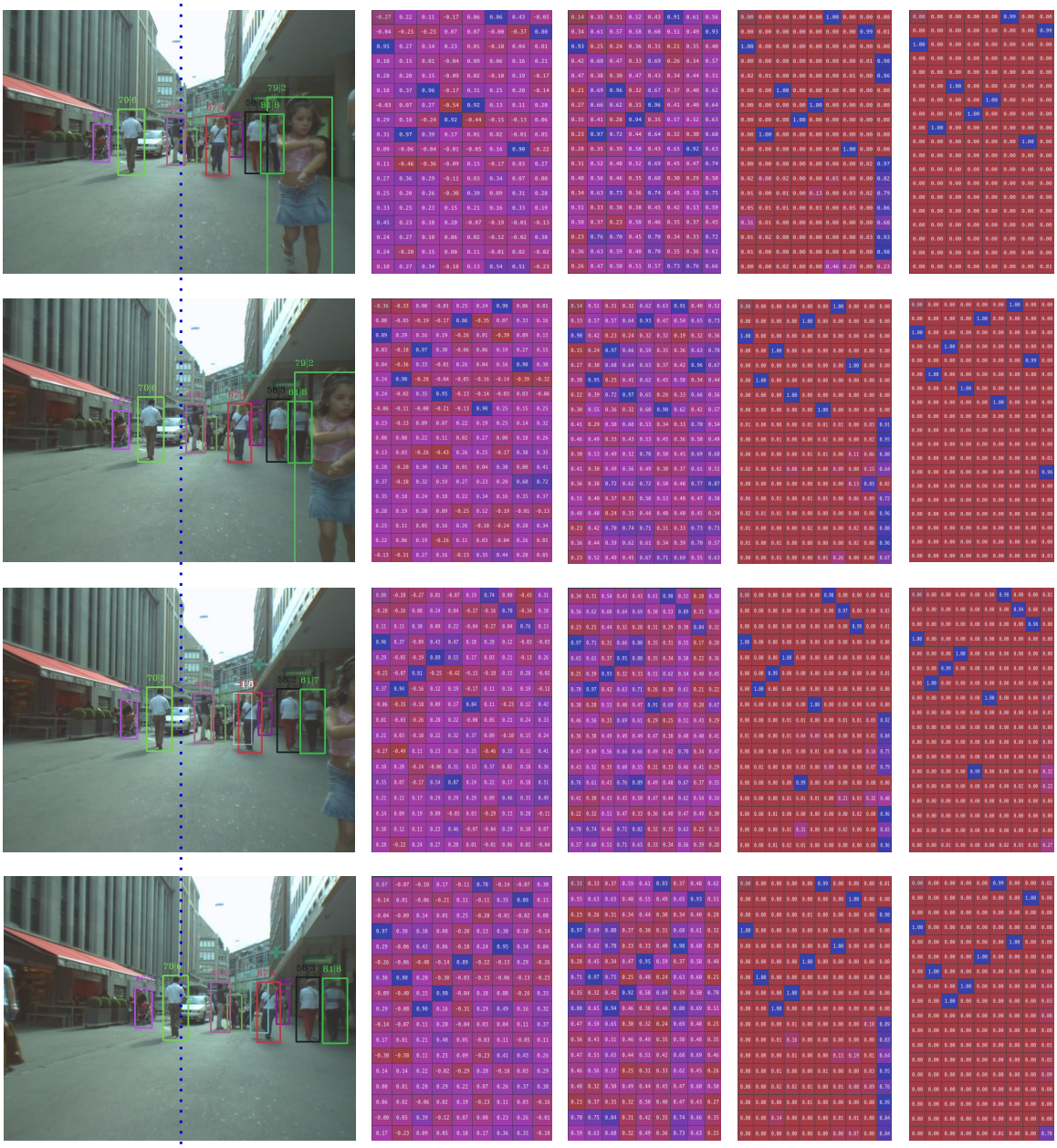


Figure 4.8 Illustrate the association results of fast-moving scenes.

two decision matrices illustrate clear associations that could be used to make association decisions.

4.6 Conclusion

This chapter discussed the application of the re-identification feature for data association. Unlike other state-of-the-art tracking methods that deemed the re-identification feature a complementary means to improve spatial-scale association accuracy, we re-evaluate the importance of the re-identification feature based on our previous proposed person search network. We conduct experiments to prove that re-identification features are more robust than spatial-scale bounding box association across all tracking scenes and all frame rate videos. We figure out that the essential problem that hinders the application of the re-identification feature is its lack of significance in the association problem. We propose an online-learning-scene-adaptive converter that online convert the re-identification features to association features. Our experiments proved that the association features could provide sufficient significant similarities between positive matches and negative matches; it is able to make reliable association decisions independently of the spatial-scale cues.

CHAPTER 5

Scene-Adaptive Detection and the Global Architecture of Multiple Object Tracker

5.1 Motivation

As aforementioned, MOT is facing two critical problems; they are object detection and data association. We detailed discussed the data association problem in chapter 4 and proposed an online-learning approach to convert the data association problem into feature space to improve the association algorithm's robustness in some challenging scenes. We argue that it is inappropriate to solve the association problem in a specific scene by pre-training a general-purpose feature extractor since each tracking scenario has its uniqueness, and the performance of trackers generally highly depends on the training datasets. Trackers can achieve corresponding results only when the real-world application scene meets the description of benchmark datasets. However, the scenarios depicted by datasets are ofter minimal. When a real-world tracking scene differs from the benchmark dataset's coverage field, detection problems like mis-detection and localization noise will greatly affect the tracking performance of non-dataset-covered applications. Seeking a way to free the tracking problem from the constraints of datasets is the main problem to be discussed in this chapter. Based on the detection and re-identification network proposed in chapter 3 and the scene-adaptive transformer proposed in chapter 4, this chapter proposes the last component of our end-to-end tracking system. The new component is

called scene-adaptive detection.

Recall rate $recall = TP/(TP + FN)$ is often used to evaluate the detection performance, where TP is true positive (true detection), FN is false negative (false detection). A detection network generally sets a hyper-parameter τ_{det} to control the recall rate. In a practical application scene, the boundary between true positive and false negative is not significant. A low threshold will increase the number of TP and FN. A high threshold will reduce the number of FN and TP. When setting a low threshold, there will be a large number of detection, the computational burden of the detection network and re-identification network will increase. Many false negative detections have Spatio-temporal continuity and are prone to be confirmed as tracks by the tracker. Thus, normally, we will set a larger threshold to avoid/decrease these false tracks. However, a larger threshold is prone to bring mis-detection that degradation the tracking continuity.

Considering the mis-detection problem, the idea of our proposed scene-adaptive detection is to feedback the tracking results to the detection network to enhance the detection confidence for specific objects.

5.2 A Breif Introduction of CenterNet

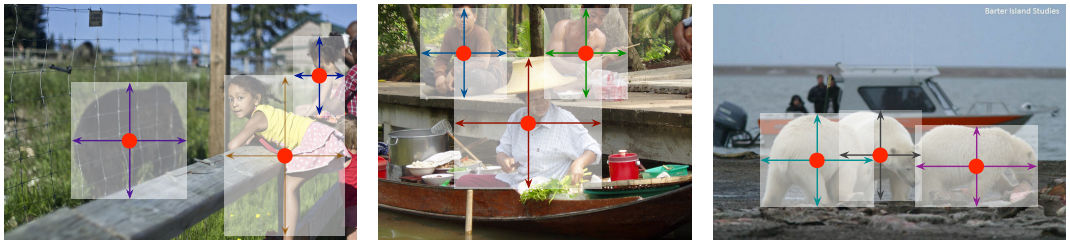


Figure 5.1 CenterNet models an object as the center point of its bounding box. The bounding box size and other object properties are inferred from the keypoint feature at the center[16].

The implementation of network architecture and loss function in different detection networks are very different. CenterNet [16] is an anchor-free implemented detection network. The loss function design is more concise and easier to realize than anchor-based methods like YOLO or MRCNN. More importantly, CenterNet regards the objects' center as key points to locate the objects' position; this could adapt the tracking task. This thesis chooses CenterNet as the underlying detection network; the implementation of the tracker is based on our previous proposed person search network—CenterNet-MDnet. In order to clarify our implementation, we give a brief introduction to CenterNet.

The head of CenterNet outputs three maps, including a keypoint heat map I^{hm} indicating the center position of objects, a regression map I^{wh} that regresses the shape (height and width) of the objects, and a regression map I^{reg} that regresses the objects' center; it compensates for quantization errors and improves positioning accuracy.

Let $I \in \mathbb{R}^{W \times H \times 3}$ be an input image with width W and height H. CenterNet aims to generate a keypoint heatmap $\hat{I}^{hm} \in [0,1]^{\frac{W}{4} \times \frac{H}{4} \times C}$, where C is the number of keypoint types; in our implementation, $C = 1$, only 'person' is considered. In the training period, the ground truth heatmap is modeled by multiple Gaussian kernels $g_{x,y} = \exp\left(-\frac{(x-\bar{p}_x)^2+(y-\bar{p}_y)^2}{2\sigma_p^2}\right)$, where σ_p is an object size-adaptive standard deviation (the size of bounding box determines the value of σ), $(\bar{p}_x, \bar{p}_y)$ is the center position of object on the heatmap, note that $(\bar{p}_x, \bar{p}_y)$ is quantized. The quantization error is modeled by I^{reg}. A prediction $\hat{I}^{hm}_{x,y} = 1$ corresponds to a detected keypoint, while $\hat{I}^{hm}_{x,y} = 0$ is the background. The objective of the training is a penalty-reduced pixel-wise logistic regression by using focal loss [100]:

$$loss^{hm} = \frac{-1}{N} \sum_{x,y} \begin{cases} (1-(\hat{I}^{hm}_{x,y})^{\alpha}) log(\hat{I}^{hm}_{x,y}) & if\ I^{hm}_{x,y} = 1 \\ (1-I^{hm^{\beta}}_{x,y})(\hat{I}^{hm}_{x,y})^{\alpha} log(1-\hat{I}^{hm}_{x,y}) & otherwise \end{cases} \tag{5.1}$$

where α and β are hyper-parameters of the focal loss and N is the number of keypoints on image I (the number of ground truth bounding boxes).

The prediction of $\hat{I}^{hm}$ is critically important in CenterNet; its local peaks are used to locate the object center. The intensity of the peak indicates the detection confidence of the objects. CenterNet will set a threshold $\tau_{det} \in [0,1]$ to filter out low confidence objects. This may results in mis-detection for small-size objects. We can infer from the equation 5.1 that the Gaussian kernel radius is small for small-size objects. The suppression of negative sample points around the center of the target is relatively large. Small-size targets naturally have much lower detection confidence than large-size objects that prone to mis-detection and causing tracking discontinuity and ID Switch. In response to this problem, this chapter proposes a scene adaptive method that uses the tracking results to online optimize the detector's detection confidence for specific objects.

5.3 Detection Confirmation

In traditional tracking-by-detection paradigm, object detection and tracking are regarded as two independent modules. The tracking/association module passively receives the detection results without feedback. Unpredictable detection errors may occur in different application scenes, such as object mis-detection and object localization noise. These detection flaws are difficult to be handled by tracker. On the other hand, detectors are doing framewise detection; the quality of detection results only depends on the current

frame and lacks the usage of video context constraints. Multiple object tracking algorithms connect the Spatio-temporal trajectory of targets in a continuous frame sequence; the tracking filter constrains objects' spatial position to refine the localization accuracy and could confirm the existence of true positive objects. A natural idea is whether we can use the tracking results to optimize the detection network.

To this end, this chapter proposes online feedback the tracking results to detection network to enhance the detection confidence of confirmed detections. Assuming that at frame t, the detector detects a set of detection $\mathbb{D}_t = \{det^1, det^2, ..., det^n\}$. Using the association algorithm proposed in chapter 4, part of detections are associated with tracks. If the corresponding tracks are confirmed to be true tracks, the associated detection could be confirmed to be true positive. We feedback this information to the detection network and 'ask' the detector to give higher detection confidence to these detections.

5.3.1 Trajectory Confirmation and High Quality Detection Selection

Due to false-negative detections, a tracker will inevitably generate false-tracks. If all tracked detections selected as positive samples, the false-negative detection will be enhanced. The enhanced false-negative detection will, in turn, further improve the tracking for false-alarm objects. This mutual influence of tracker and detector is dangerous that will cause tracking and detection diverging. Therefore, it is critically important to select positive tracks. We use the following two limitations to screen out true positive tracks and detections.

- $track_t = \{node_{t-n+1}, node_{t-n+2}, ..., node_t\}$ is a track updated in frame t and consists of n nodes from frame $t-n+1$ to frame t. In order to confirm this trajectory to be a positive trajectory, the value of n is required to be greater than a specific value, i.e., 5. And requires at least one node whose detection confidence higher than a threshold $\tau_n \in [0, 1]$. In principle, $\tau_n > \tau_{det}$.
- The confirmed trajectory must contain at least one node with the width and height greater than a specific value. Assuming that the target's width and height are (w, h), and the corresponding Gaussian kernel radius is r. To confirm the trajectory, r need to larger than a threshold $\tau_r \in (1, +\infty)$.

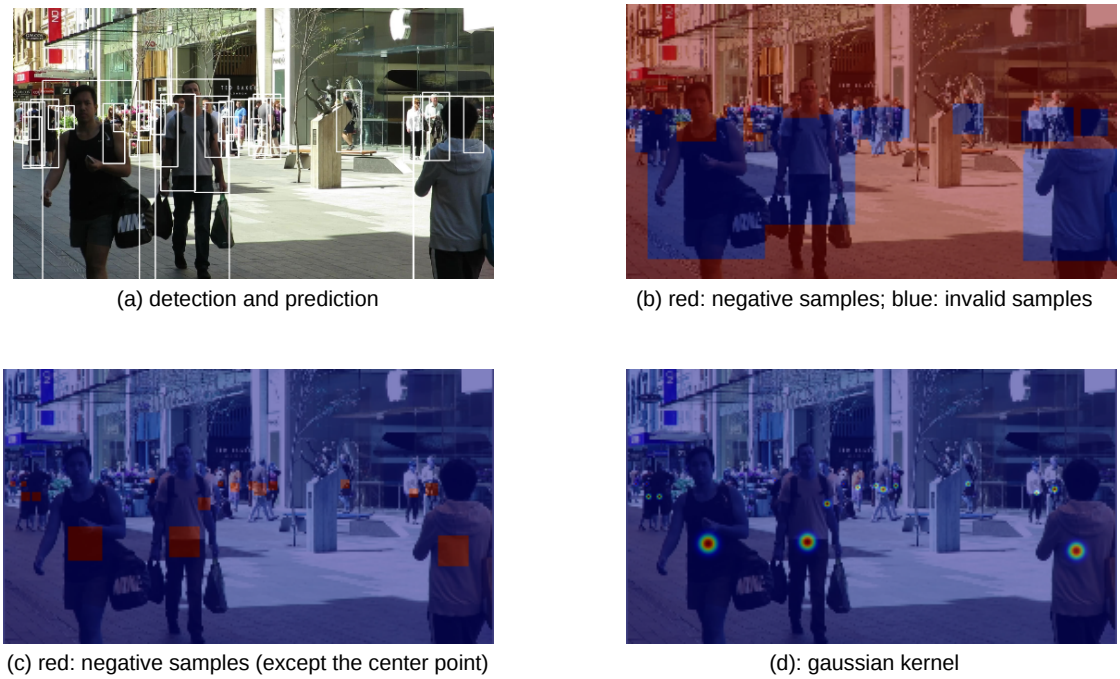
Through these two preliminary screening steps, false-negative detection or false tracks will be eliminated, and the high-quality targets will be screened out. It is worth noting that the confirmation of detection is based on the historical information of the track rather than the current detection. With forward calculation of the tracking process, the detection confidence of the tracked target may be lower than the detection threshold of the detector,

resulting in object mis-detection. If this situation can be fed back to the detection network in time, the detector could enhance the target and increase the continuity of tracking. In principle, this process also enhances the detection of other untracked targets.

5.3.2 The Design of Heatmap and Loss Function

We intend to feedback the tracking results to the detection network to enhance the detector's adaptability to the current scene. CenterNet has three outputs, including a heatmap $\hat{I}^{hm}$ to predict the center of objects, a height and width regression map $\hat{I}^{wh}$ to predict the height and weight of objects, and a position regression map $\hat{I}^{reg}$ to improve the localization accuracy. As afore discussed, Due to the nonlinearity of object motion, the Kalman filter cannot reduce the detection noise and may introduce non-linear noise, the filtering results of the tracker are not suitable to optimize the height, width, and position of detection. Instead, feeding back the heatmap to the detection network and updating the detection network does not essentially destroy the detector's performance if we find sufficient positive samples and negative samples.

Positive Samples and Negative Samples



(a) detection and prediction

(b) red: negative samples; blue: invalid samples

(c) red: negative samples (except the center point)

(d): gaussian kernel

Figure 5.2 Positive Samples and Negative Samples.

Fig 5.2 shows the training samples.

— Image *(a)*: the white bounding boxes are detections generated by the detection network and the predictions generated by tracker.
— Image *(b)*: the red regions out of the coverage of all bounding boxes are confirmed to be negative samples. The blue regions are unconfirmed (invalid) samples and will not be used as training samples.
— Image *(c)*: the detection bounding boxes confirmed by tracker to be true detection will mark the corresponding region as negative (except the center point, which is a positive sample). The negative samples will be used to transform the invalid region of Image *(b)*. (invalid samples transformed to negative sample and positive sample).
— Image *(d)*: the confirmed detection are represented by Gaussian kernels.

Online learning is different from training on datasets. A labeled dataset will mark out all positive samples. The unlabeled regions automatically become negative samples. In our case, the center keypoint of bounding boxes are positive samples. The small effective areas except for those center point key points of confirmed detection are deemed negative. Besides, the regions outside all detection and prediction bounding boxes are negative samples. In this case, we could get enough positive samples and negative samples.

We denote the heatmap as $I^{hm}_{x,y}$. Then the effective area of the heatmap could be written as:

$$I^{e}_{x,y} = \begin{cases} 1 & if\ I_{x,y} > 0\ and\ I_{x,y} valid \\ 0 & otherwise \end{cases} \tag{5.2}$$

where x, y is the indices of elements. The pixels outside the invalid regions are valid pixels.

The Design of the Loss Function

In the original implementation of CenterNet, all pixels on the heatmap are negative samples except for the center point of the Gaussian kernel. In our proposed method, only the effective areas of the heatmap are taken into consideration. The feedback loss function is:

$$loss^{hm} = -\frac{1}{N}\sum_{x,y}\begin{cases} (1-(\hat{I}^{hm}_{x,y})^{\alpha})log(\hat{I}^{hm}_{x,y}) & if\ I^{hm}_{x,y} = 1 \\ (1-I^{hm^{\beta}}_{x,y})(\hat{I}^{hm}_{x,y})^{\alpha} log(1-\hat{I}^{hm}_{x,y}) I^{e}_{x,y} & otherwise \end{cases} \tag{5.3}$$

where $\hat{I}^{hm}_{x,y}$ is the predicted heatmap generated by detector. $I^{hm}_{x,y}$ is the ground truth heatmap generated by tracker. And $I^{e}_{x,y}$ is the effective area of the heatmap.

5.4 The Global Archetecture of Scene-Adaptive Tracking System

We split CenterNet into two parts—$CenterNet^{-}$ and $conv^{hm} = \{\mathbf{K}^{hm}, \mathbf{B}^{hm}\}$. $conv^{hm}$ is the last convolutional layer that outputs $\hat{I}^{hm}_{x,y}$. K^{hm} and B^{hm} are the corresponding convolutional kernel and bias vector. The prediction of heatmap could be written as:

$$\hat{I}^{hm}_{x,y} = conv^{hm}(CenterNet^{-}(I)) \tag{5.4}$$

where I is the input image. The global architecture of our proposed scene-adaptive

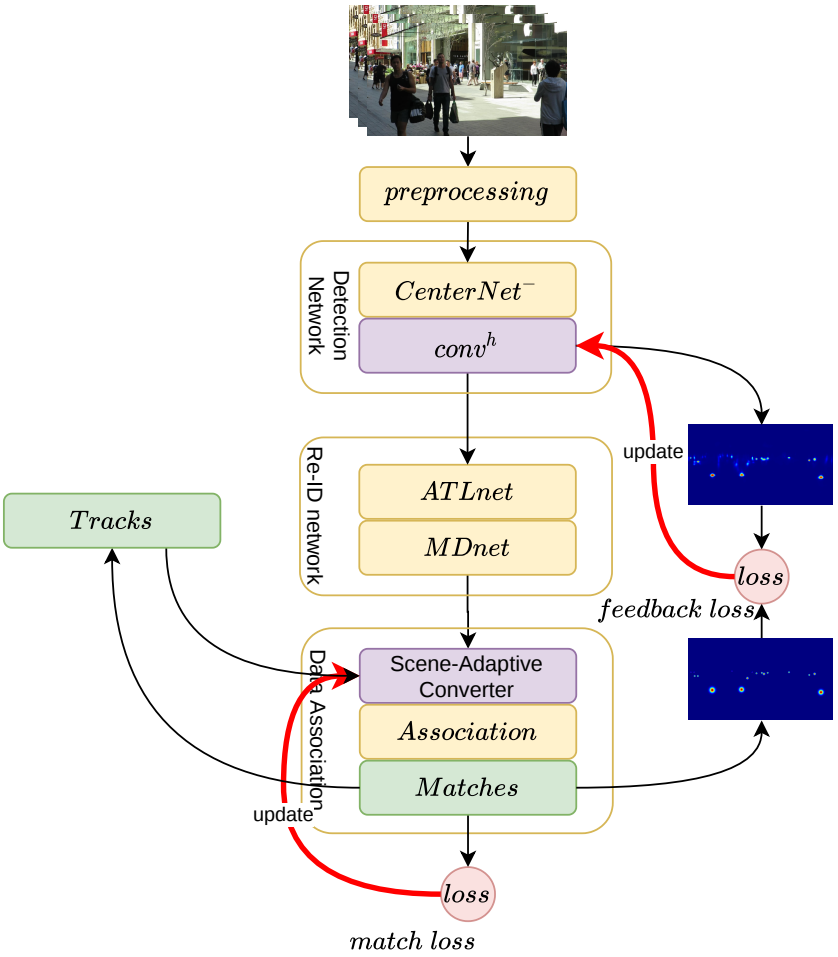


Figure 5.3 The global architecture of scene-adaptive tracking system.

tracking system is shown in Fig. 5.3. The overall processes are listed as follow:

1 **Preprocessing**: The input of the tracking system is images; the images are first resized to a specific shape to fit the input shape of the detection network.

2 **Object Detection**: The resized image computed by the detection network to generate detection bounding boxes. By the way, the heatmap $\hat{I}^{hm}_{x,y}$ predicted by the convolutional layer $conv^{hm}$ is also output.

3 **Re-identification Feature Extraction**: The re-identification sub-network $ATLnet$ and $MDnet$ extracts appearance features for each detected bounding box.

4 **Scene-Adaptive Converter**: The converter was performed to convert the detection's re-identification features and the tracks' re-identification features to association features.

5 **Cascade Association**: Using the association algorithm proposed in chapter 4 to solve the association problem. The association consists of two cascade steps. The first step is to get $matches^{reid}$ from solving the two $Decision-Matrix$ 4.24. The second step is to get $matches^{join}$ from solving the join spatial-scale and feature space association weight matrix 4.17.

6 **Updating Tracks**: Using matching results to update the track list.

7 **Updating Converter**: Based on the matches, the match loss 4.26 back propagates to the scene-adaptive converter 4.27 to update the weights.

7 **Creating Heatmap**: Tracks that confirmed as true positive will be used to create a Gaussian kernel heatmap $I^{hm}_{x,y}$.

8 **Updating Detection Network**: Get the feedback loss 5.3 and back propogates it to the detection network to update the weights of $conv^{hm}$ by:

$$\begin{aligned} \mathbf{K}^{hm}_{t|t+1} &= \mathbf{K}^{hm}_{t-1|t} + \eta \frac{\delta_{loss^{hm}}}{\delta_{\mathbf{K}^{hm}_{t-1|t}}} \\ \mathbf{B}^{hm}_{t|t+1} &= \mathbf{B}^{hm}_{t-1|t} + \eta \frac{\delta_{loss^{hm}}}{\delta_{\mathbf{B}^{hm}_{t-1|t}}} \end{aligned} \tag{5.5}$$

5.4.1 The updating of track list

The updating of the track list could be modeled as a state machine updating problem; see Figure 5.4. A track has 4 states as following:

— **New**: Detection that unassociated by any existed tracks, the spatial-scale distance and feature space distance sufficient large, and the detection confidence larger than a threshold τ_{new} will be confirmed as a new track.

— **Tracked**: A track associated with a current frame's detection is marked as a tracked track.

— **Lost**: A track that failed to associate a new detection is marked as a lost track.

— **Removed**: A track removed from the track list and will not be used to associate with new detection.

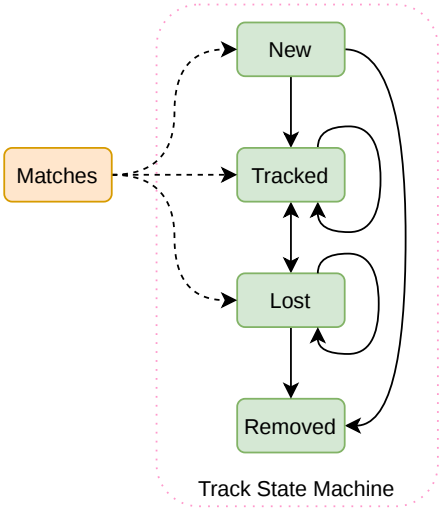


Figure 5.4 Track State Machine and Association Pipeline.

The updating of the track state machine could be described as follow:

- **New→Tracked**: A **New** track associate with a new detection. The state of the track upgrade to **Tracked**
- **New→Removed**: A **New** track failed to associate a detection, *remove* the track immediately.
- **Tracked→Tracked**: A **Tracked** associate a new detection. The track's state remains unchanged.
- **Tracked→Lost**: A **Tracked** track failed to associate with a new detection; the state of the track degrade to **Lost**
- **Lost→Tracked**: A **Lost** track associate a detection, and the detection confidence larger than τ_{new}; the state of the track upgrade to **Tracked**
- **Lost→Lost**: A **Lost** track failed to associate with a new detection. The state of the track's state remains unchanged.
- **Lost→Removed**: A **Lost** track unassociate with a new detection for multiple frames,i.e., 30 frames; *remove* the track from track list.

5.5 Experiments and Evaluation

In this section, we conduct experiments to validate the proposed tracking system. In 5.5.1 we will get the detailed parameter setting of the scene-adaptive detection component and the updating of the track list. In 5.5.2 we will visualize the results of the scene-adaptive detection components. In 5.5.3 we will visualize the tracking results. In 5.5.4 we will give the tracking performance by using MOTchallenge benchmark.

5.5.1 Implementation Details

As shown in Fig. 5.3, our end-to-end tracking system consists of a person search network—CenterNet-MDnet, a association module, and a track list updating module. The parameter setting of CenterNet-MDnet 3 and the association algorithm 4 has been discuessed in the previous chapters. Here, we give the parameter setting for the scene-adaptive detection component and the updating of track list. The detection confidence threshold τ_{det} is set to 0.3. The new track confirmation threshold τ_{new} is set to 0.4. The learning rate η of equation 5.5 is set to 0.002. The track confirmation threshold τ_n is set to 0.7.

5.5.2 Visualization of the Detection Results of Scene-Adaptive Detection

Fig. 5.5 shows the results of using scene-adaptive detection. MOT17-06 was recorded by a hand-held mobile camera with large target scale and number variations. We employ it to verify the effectiveness of our proposed methods. MOT17-06 contains 1194 frames; we choose frames 137, 827, and 966 that well cover the whole videos to demonstrate the detection results. The left-most column of Fig. 5.5 is the tracking results and the corresponding heatmaps. It could be seen that not all tracks are used to generate the heatmap, instead, we choose high-quality and long-standing tracking objects to generate the objective heatmap. The middle column are the detection results and heatmaps of the original implementation of CenterNet. The right column are the detection results and heatmaps by using our scene-adaptive detection. For comparison, our method effectively enhances the detection of low-confidence objects, especially small objects. It worth noting that, in the whole tracking process, the feeding back gradients did not destroy the detection network. The background of the output heat map is relatively clean (even less clutters than the original implementation). The number of false-negative detections is at the same level as the original detection network.

Our scene-adaptive detection method can effectively enhance the detection ability of the detection network, which proves that the refine method of feedback tracking results to the detection network is effective.

The object detection network's performance has a decisive influence on the multiple object tracking algorithm's performances. General-purpose detection networks usually cannot obtain optimal detection results in a specific tracking environment or requires hand-designed parameters to fetch a good recall rate. We argue that if the detection network can adaptively optimize for each specific scene, its detection capability still has room for improvement.

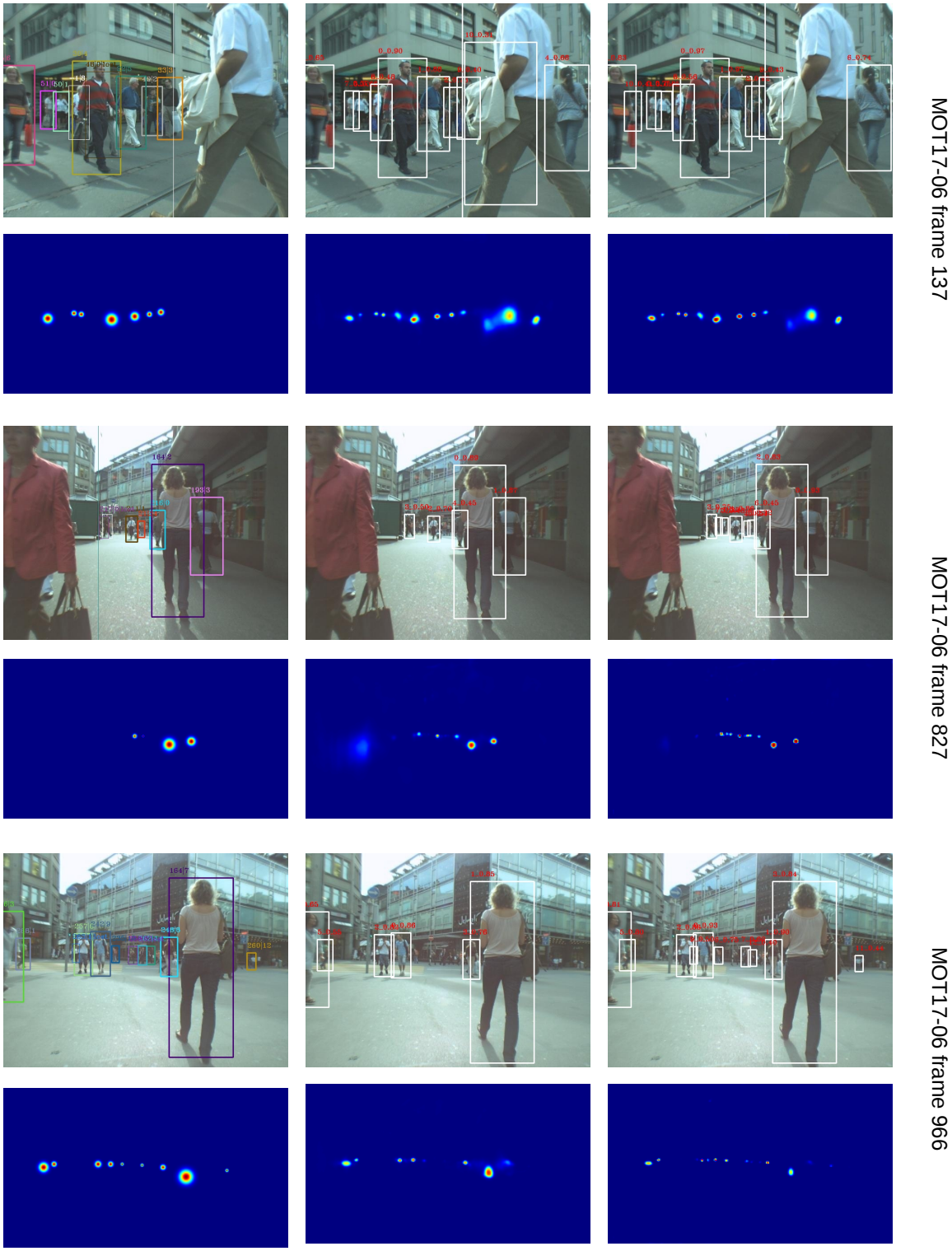


Figure 5.5 Visualization of Scene-Adaptive Detection.

5.5.3 Visualization of the Tracking Results

Visualization of the Tracking of Intersection Occlusion Scene

Figure 5.6 Visualization of the Tracking of Intersection Occlusion Scene.

Fig. 5.6 shows the tracking of intersection occlusion scene. Bounding boxes with the same color indicate the same identity. In this challenging tracking scene, the frequently track intersection causing frequent object occlusion and mis-detection is not robust to get matches via spatial-scale distance association. Some state-of-the-art trackers like [68, 69] using bounding box as the main association cues will generate a large number of trajectory segments or ID switches due to lack of spatial-scale association constraints. Our proposed tracking system, mainly using re-identification features for the association, could better adapt these difficult tracking scenes.

Figure 5.7 Visualization of the Tracking of Crowded Scene.

Visualization of the Tracking of Crowded Scene

Fig. 5.7 shows the tracking results of crowded scene. MOT17-03 is a static camera crowded tracking scene with around 70 person identities per frame, and in some region of the tracking scene, the pedestrians are frequently mutual occluded. In this type of scenarios, due to the small spatial-scale distance and detection noises, bounding box association is likely to cause association errors, which leads to ID switches. Our method, as shown in 4.7, due to the application of scene-adaptive converter, most association matches are solved in the association feature space. This give our method a large improvement on the ID-Switch problem.

5.5.4 Evaluation of the Tracking Performance

In order to comprehensively compare the performance of our proposed tracking system, we evaluated the performance on the public dataset—MOT17.

Table 5.1 MOT17 Testing Sequences

MOT17 Testing Sequences							
Name	FPS	Resolution	Frames	Tracks	Boxes	Viewpoint	Camera
MOT17-01	30	1920x1080	450	23	6395	medium	static
MOT17-03	30	1920x1080	1500	148	104556	high	static
MOT17-06	14	640x480	1194	217	11538	medium	moving
MOT17-07	30	1920x1080	500	55	16322	medium	moving
MOT17-08	30	1920x1080	625	63	16737	medium	static
MOT17-12	30	1920x1080	900	94	8295	medium	moving
MOT17-14	25	1920x1080	750	230	18483	high	moving
Total			5919	830	182326		

Table 5.2 Overall Evaluation Results

Tracker	**MOTA↑**	**IDF1↑**	**MT↑**	**ML↓**	**ID SW↓**	**Frag↓**
FairMOT[68]	73.7	72.3	1017	408	3303	8073
MAT[57]	69.5	63.1	1032	444	2844	**3726**
CTTracker[101]	67.8	64.7	816	579	3039	6102
CSTrack[102]	**74.9**	**72.6**	978	411	3567	7668
SATracker(Ours)	73.12	72.39	1023	375	1569	4716
SATracker+(Ours)	74.25	72.25	**1095**	**294**	**1416**	4257

Testing Datasets

MOT17 contains 7 testing video sequences for evaluating the tracker's performance; the information of videos are shown in Table 5.1. The scenes of MOT17 vary significantly in terms of background, illumination conditions, camera viewpoints, and camera motion. These variations make MOT17 very challenging. It worth noting that the labels and ground truth bounding boxes of the testing videos are not available publicly. The evaluation is performed on the MOTchallenge benchmark [1].

Evaluation Results

Our tracker tracks the testing videos on the local machine; the tracking results are recorded in *.txt* formatted files and uploaded to the host server for evaluation. The detailed evaluation metrics of the multi-object tracker has been given in 2.4. The overall evaluation results were shown in Table. 5.2. We summarized the evaluation results of current paper-

1. https://motchallenge.net

available online tracking algorithms on the leader board of the MOTchallenge server. As discussed in previous chapters, the tracker's performance, especially the MOTA, is mainly determined by the detection network. Our proposed method uses the same (same network architecture and same network weights) detection network as FariMOT [68], the MOTA is close. To evaluate the function of each proposed component, we designed two implementations. The first is Scene-Adaptive Tracker, abbreviates as SATracker. It consists of two components, including a person search network—CenterNet-MDnet for pedestrian detection and re-identification feature representation. And the scene-adaptive converter for data association. The second implementation is Scene-Adaptive Tracker Plus, abbreviates as SATracker+, an upgrade of SATracker that composes an additional scene-adaptive detection module. The design of the scene-adaptive association module is to convert the data association problem from spatial-scale association to feature space association to improve the tracker's tracking capability for difficult scenes, i.e., fast-moving scenes. It could be seen from Table. 5.2 that the tracker SATracker has greatly improved the tracking continuity while maintaining the MOTA to 73.12%. When comparing with FairMOT, the IDSW cases decreased to half of it (3303 to 1569), and the track fragments also decreased for a large number (8073 to 4716). SATracker+ is equipped with a scene-adaptive detection module designed to enhance the detection of weak objects that further improve the tracking continuity. The mostly lost (ML) tracks decreases from 375 to 294.

5.6 Conclusion

This chapter proposes a scene-adaptive detection module. The idea is to use the tracking results to enhance the detector's confidence to detect weak positive objects. We propose the global architecture of the tracking system and detailed discussed the updating of the tracklist. Based on the proposed tracking system, we get tracking results on the challenging tracking dataset—MOT17. The experimental results show that our scene-adaptive association and scene-adaptive detection modules are effective in improving the tracker's performance. The application of re-identification feature association and detection enhancement greatly improve the track's tracking continuity. Throughout the design, we emphasized the importance of feeding back the information. The association results fed back to the scene-adaptive converter to update the weights of it. The tracking results fed back to the detection network to update the weights of the last convolutional layer. We believe that 'learning' should not only exist in the training phase. Online-learning the current scene is critically important to improve the tracking performance.

CHAPTER 6

Conclusions and Scope for Future Work

6.1 Conclusions

Implement an end-to-end multiple object tracking system is a challenging task due to the need to comprehensively take object detection, appearance feature representation, and data association into consideration. We have proposed novel approaches for these three aspects and designed a complete, robust tracking system. The final tracker proposed in this thesis has the following innovative properties:

- For pedestrian appearance feature representation, we propose a deep-frozen transfer learning framework for person search to extracts re-identification features straightly from a pre-trained detection network.
- For the data association problem, we propose a scene-adaptive data association solution to transfer the re-identification features to the association feature to improve the tracking continuity in fast-moving scenes.
- For the object detection problem, consider that the application of detection network in practical scenes unavoidable cause mis-detection problem; we propose a scene-adaptive detection module to online optimize the last convolutional layer of the detection network to generate more robust detection for true-positive detection.

We have made contributions in multiple aspects. In the design of the transfer learning framework for person search, we proved that the feature of pre-trained detection well

adapts to the task of re-identification. This conclusion allowed us to design a shallow additional re-identification module to extract high-performance re-identification features. The experimental results show that our solution outperforms all other person search solutions by a large margin. This critically important to ease the design of integrating object detection network and re-identification network.

We then discussed the re-identification feature application problem. For the first time, we give a brief introduction to the re-identification feature space, point out that a well-trained re-identification technically contains redundancies for data association, following traditional solutions that directly apply the re-identification feature for data association will never successed. We propose to use an online-learning solution to deal with the re-identification association problem. This is a heuristic exploration, which provides a new solution to the object association algorithm.

Last, we propose to interact between the tracker and the detector to selectively enhance the detection of true-positive targets. We believe that the information exchange between the detector and the tracker plays a vital role in improving tracking performance.

6.2 Future Work

We have improved the overall tracking algorithm framework in many aspects, and believe that these methods still have room for improvement.

— The definition of re-identification features is fuzzy, and the composition of re-identification features is still a mystery. Although our solution can achieve high pedestrian search accuracy in datasets, due to our lack of visualization research on the re-identification network and lack of quantitative analysis on feature composition, we are still designing the network structure intuitively, lacking a theoretical foundation. In future research, we believe that visualization research of the re-identification network can improve the interpretability of feature vectors and further improve the capability of the re-identification feature.

— Our proposed scene-adaptive methods are pretty superficial. They should have more in-depth research. Like our human vision systems, we believe that adaptive learning methods are the key to applying fixed pre-trained knowledge to practical scenarios. Maybe we can develop more dynamic learning components to adapt to more practical scenes. Such as illumination adaptive module.

In addition to these, the detection network has a decisive influence on the tracker. The existing detectors more or less have overfitting problems. In our further research work, we should integrate all methods, including redesigning/retrain the detection network to

achieve better tracking performance.

Bibliography

[1] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3415–3424.

[2] 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.

[3] Q. Pan, X. Ye, F. Yang, and H. Zhang, "Generalized probability data association algorithm," *Idea*, vol. 529, no. 103, p. 150, 2005.

[4] S. Hamid Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid, "Joint probabilistic data association revisited," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3047–3055.

[5] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4696–4704.

[6] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[7] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[9] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[10] ——, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[12] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[15] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[16] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[21] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[23] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[24] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 390–391.

[25] D. Misra, "Mish: A self regularized non-monotonic neural activation function," *arXiv preprint arXiv:1908.08681*, 2019.

[26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020.

[27] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4516–4524.

[28] S. Bak, S. Zaidenberg, B. Boulay, and F. Bremond, "Improving person re-identification by viewpoint cues," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2014, pp. 175–180.

[29] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3765–3773.

[30] Y. Huang, Z.-J. Zha, X. Fu, and W. Zhang, "Illumination-invariant person re-identification," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 365–373.

[31] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1354–1362.

[32] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.

[33] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1288–1296.

[34] 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 *Proceedings of the iEEE conference on computer vision and pattern recognition*, 2016, pp. 1335–1344.

[35] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in

Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 480–496.

[36] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 384–393.

[37] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[38] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "Alignedreid++: Dynamically matching local information for person re-identification," *Pattern Recognition*, vol. 94, pp. 53–61, 2019.

[39] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[40] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[41] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[42] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–419.

[43] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision*. Springer, 2016, pp. 475–491.

[44] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3702–3712.

[45] J. Li, F. Liang, Y. Li, and W.-S. Zheng, "Fast person search pipeline," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1114–1119.

[46] X. Lan, X. Zhu, and S. Gong, "Person search by multi-scale matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 536–552.

[47] Y. Lu, Z. Hong, B. Liu, W. Li, and N. Yu, "Dhff: Robust multi-scale person search by dynamic hierarchical feature fusion," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3935–3939.

[48] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream cnn model," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[49] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 937–940.

[50] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng, "Ian: the individual aggregation network for person search," *Pattern Recognition*, vol. 87, pp. 332–340, 2019.

[51] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan, "Neural person search machines," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 493–501.

[52] H. Liu, W. Shi, W. Huang, and Q. Guan, "A discriminatively learned feature embedding based on multi-loss fusion for person search," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1668–1672.

[53] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2158–2167.

[54] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 811–820.

[55] A. Loesch, J. Rabarisoa, and R. Audigier, "End-to-end person search sequentially trained on aggregated dataset," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4574–4578.

[56] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[57] S. Han, P. Huang, H. Wang, E. Yu, D. Liu, X. Pan, and J. Zhao, "Mat: Motion-aware multi-object tracking," *arXiv preprint arXiv:2009.04794*, 2020.

[58] Y. Xu, A. Osep, Y. Ban, R. Horaud, L. Leal-Taixé, and X. Alameda-Pineda, "How to train your deep multi-object tracker," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6787–6796.

[59] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multi-target tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 58–72, 2013.

[60] S. Oh, S. Russell, and S. Sastry, "Markov chain monte carlo data association for multi-target tracking," *IEEE Transactions on Automatic Control*, vol. 54, no. 3, pp. 481–497, 2009.

[61] S. Ba, X. Alameda-Pineda, A. Xompero, and R. Horaud, "An on-line variational bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, 2016.

[62] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[63] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 941–951.

[64] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, 2008.

[65] 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*. Springer, 2016, pp. 17–35.

[66] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2953–2960.

[67] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[68] Y. Zhan, C. Wang, X. Wang, W. Zeng, and W. Liu, "A simple baseline for multi-object tracking," *arXiv preprint arXiv:2004.01888*, 2020.

[69] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi-object tracking," *arXiv preprint arXiv:1909.12605*, 2019.

[70] X. Wan, J. Wang, and S. Zhou, "An online and flexible multi-object tracking framework using long short-term memory," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1230–1238.

[71] W.-J. Tsai, Z.-J. Huang, and C.-E. Chung, "Joint detection, re-identification, and lstm in multi-object tracking," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[72] Y. Lu, C. Lu, and C.-K. Tang, "Online video object detection using association lstm," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2344–2352.

[73] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[74] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6036–6046.

[75] Y. Yang, S. Liao, Z. Lei, and S. Z. Li, "Large scale similarity learning using similar pairs for person verification," in *Thirtieth AAAI conference on artificial intelligence*, 2016.

[76] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[77] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[78] 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*. Springer, 2016, pp. 17–35.

[79] K. Islam, "Person search: New paradigm of person re-identification: A survey and outlook of recent works," *Image and Vision Computing*, vol. 101, p. 103970, 2020.

[80] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2359–2367.

[81] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1367–1376.

[82] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 2661–2671.

[83] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

[84] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[85] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and

fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[86] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, 2020.

[87] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3691–3701.

[88] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[89] W. Sun, F. Liu, and W. Xu, "Unlabeled samples generated by gan improve the person re-identification baseline," in *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, 2019, pp. 117–123.

[90] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 2138–2147.

[91] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8933–8940.

[92] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[93] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[94] 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*. Springer, 2014, pp. 740–755.

[95] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[96] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[97] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[98] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[99] S. Sun, N. Akhtar, H. Song, A. S. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[100] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[101] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," *ECCV*, 2020.

[102] C. Liang, Z. Zhang, Y. Lu, X. Zhou, B. Li, X. Ye, and J. Zou, "Rethinking the competition between detection and reid in multi-object tracking," *arXiv preprint arXiv:2010.12138*, 2020.

Titre de la thèse:

Apprentissage profond de bout-en-bout pour la ré-identification et le suivi de personnes

septembre, 2021
HU Ronghua

French Summary

Introduction

Le suivi d'objets multiples consiste à traiter de manière entièrement automatisée des séquences d'images et de vidéos pour la localisation objets d'intérêt et d'estimation de leurs trajectoires de mouvement spatio-temporel. Grâce à une base de données techniques et aux technologies d'apprentissage profond, le suivi et la détection d'objets multiples sont utilisés dans le domaine de la sûreté et de la sécurité. En raison du changement d'apparence des personnes, de leurs mouvements non linéaires et de leurs occlusions mutuelles dans des scènes de foule et de mobilité, le suivi reste extrêmement complexe. Un système de suivi complet et robuste se compose d'un détecteur pour la détection sémantique, d'un réseau de réidentification pour la représentation de l'apparence des piétons, et d'un module d'association pour la maintenance et la mise à jour des trajectoires. Dans cette thèse on vise à intégrer les modules de détection en utilisant des technologies d'apprentissage approfondie pour la recherche et le suivi des piétons en mouvement par : I. L'extraction de la représentation des caractéristiques d'apparence des personnes. Elle vise la localisation et la correspondance des piétons sur une galerie d'images de caméras croisées. II. Le convertisseur adaptatif de scène convertit les caractéristiques de ré-identification en caractéristiques d'association afin de prendre des décisions d'association sans la contrainte des boîtes englobantes. III. La remontée en ligne des résultats de suivi vers le réseau de détection pour améliorer la détection des cibles petites et faibles.

Chapitre 1: Contenu et Portée de la Recherche

Le détecteur d'objets fournit des informations sur l'échelle spatiale de l'objet, comme sa position, sa taille, etc. Dans les scènes de foule ou les scènes où la caméra se déplace rapidement, les informations à l'échelle spatiale ne peuvent pas fournir de contraintes discriminatoires suffisantes. Une piste peut facilement perdre son association ou se voir attribuer une mauvaise association. Dans de nombreux algorithmes de suivi de pointe, les descripteurs d'apparence des objets sont utilisés pour mesurer la distance dans l'espace des éléments afin d'améliorer la robustesse de l'association. L'intégration d'une image de piéton en 2D dans un vecteur de caractéristiques en basse dimension en 1D et la fourniture d'une discrimination suffisante est une technologie difficile mais prometteuse pour résoudre le problème de l'association. Les deux thèmes de la vision par ordinateur - la réidentification des personnes et la recherche de personnes en se concentrant sur la représentation des caractéristiques d'apparence de la caméra croisée - répondent à l'exigence de fournir une mesure de similarité de discrimination pour les problèmes de suivi difficiles, comme le suivi de la perte de piste de la caméra croisée et l'association d'objets se déplaçant rapidement. Cependant, bien que les fonctions de réidentification soient bien adaptées aux exigences du suivi, elles ont été traitées comme un moyen auxiliaire pour améliorer la robustesse de l'association à l'échelle spatiale pour les scènes de foule. Lorsque les contraintes à l'échelle spatiale ne sont pas satisfaites, le problème du suivi d'objets en mouvement rapide ne peut toujours pas être résolu correctement.

Ces dernières années, les communautés de la vision par ordinateur ont proposé de nombreux nouveaux ensembles de données standard, des points de référence et de nouvelles architectures de réseaux neuronaux profonds qui incitent les algorithmes de détection d'objets à faire des progrès remarquables dans la classification des objets et la précision de leur localisation. Les détecteurs de pointe peuvent classifier des objets de classes multiples, régresser les boîtes de délimitation de l'objet et même segmenter des instances d'objets sémantiques au niveau du pixel. Les difficultés de détection rencontrées par le passé, comme la détection d'objets d'occlusion, la détection d'objets à plusieurs échelles et l'efficacité de l'inférence des réseaux de détection, s'améliorent progressivement avec la proposition de nouveaux algorithmes. La conception des systèmes de suivi d'objets multiples entièrement automatiques varie en fonction de l'évolution des technologies de détection. Dans le cadre du paradigme du suivi par détection, la détection et le suivi d'objets sont considérés comme deux tâches successives, et les résultats de la détection ont un effet décisif sur les performances du suivi. Lorsque les scènes d'application pratique réelle varient en termes d'éclairage, de résolution et de chrominance, les différences par rapport aux ensembles de données d'entraînement standard seront inévitables pour apporter des bruits de détection et des erreurs de détection au système de suivi. En

général, les trackers reçoivent passivement les résultats de détection des détecteurs, et les problèmes de détection existants sont difficiles à résoudre correctement par les trackers, ce qui entraîne une dégradation des performances de suivi. Il existe toujours un écart entre les ensembles de données de formation standard et les applications pratiques d'inférence.

L'association de données reliant la détection de plusieurs trames à la suite pour générer des trajectoires de continuité est la technologie de base dans la conception d'un système de suivi de bout en bout. Les défis posés par l'association de données sont dus à un certain nombre de problèmes potentiels de vision artificielle : les informations de fond des objets et les informations sur l'environnement qui sont perdues par le réseau de détection déterminent que le suivi par vision artificielle ne pourrait pas atteindre les performances et la robustesse de notre système de vision humaine. L'absence d'informations sur les mouvements des cibles et des caméras dilue les contraintes à l'échelle spatiale des objets, ce qui entraîne des problèmes de suivi des pertes ou d'attribution de mauvaises associations. Aucun algorithme d'association basé sur l'apprentissage bien défini n'a été proposé jusqu'à présent. L'association de données utilisant généralement des technologies traditionnelles est modélisée comme un problème d'affectation de données. L'algorithme hongrois (HA) ou l'algorithme Kuhn-Munkres (KMA) dominent les solutions au problème d'association de données.

Pour résoudre le problème du suivi d'objets multiples par vision, la majorité des systèmes de suivi de pointe dans la littérature utilisent le paradigme du suivi par détection. Outre la détection d'objets et l'association de données, la représentation de caractéristiques d'apparence est un élément essentiel ajouté au pipeline de suivi pour une association de données robuste permettant d'adapter des scènes de suivi plus complexes, en particulier dans les réseaux de caméras non chevauchantes. Cette thèse vise à rassembler ces trois composants dans un cadre uniforme de bout en bout en utilisant des technologies d'apprentissage approfondi. L'objectif principal est de concevoir un réseau neuronal profond unique à branches multiples capable de suivre des objets multiples dans des réseaux de caméras se chevauchant ou non. À cette fin, notre travail principal contient trois contenus de recherche, chacun contribuant à l'ensemble final du système de suivi de la MOT.

— La représentation de l'apparence des piétons est un élément important du suivi des piétons par les réseaux de caméras. Elle fournit une métrique de similarité dans l'espace des caractéristiques, ce qui permet d'améliorer la robustesse de l'association des données. La représentation de l'apparence peut être interprétée comme l'intégration d'une image 2D d'un piéton dans un vecteur de caractéristiques 1D à faible dimension qui élimine les informations redondantes et conserve les informations discriminantes essentielles. La réidentification des personnes est

un sujet de vision par ordinateur largement étudié. Elle utilise des technologies d'intégration pour la représentation de l'apparence des piétons et vise à faire correspondre des images de piétons recadrées manuellement entre des requêtes et une galerie de candidats dans des systèmes de surveillance à caméras croisées. Toutefois, en raison de l'absence de module de détection des piétons, elle présente des limites d'application dans des scènes pratiques. La recherche de personnes est plutôt un nouveau sujet de vision par ordinateur proposé pour l'application de bout en bout des technologies de réidentification. Elle intègre des technologies de détection des piétons et de représentation des caractéristiques de l'apparence des personnes pour localiser et faire correspondre un piéton interrogé à une galerie d'images brutes de caméras croisées. Face à une conception plus compliquée des algorithmes de détection, la plupart des solutions de recherche de personnes de pointe n'introduisent pas de nouveaux réseaux de détection et ont tendance à utiliser des réseaux de détection plutôt simples qui se traduisent par une faible précision de détection, une faible efficacité d'inférence et une faible précision de ré-identification. Cette thèse se concentre sur l'étude de la "recherche de personnes". Nous proposons un cadre d'apprentissage de transfert surgelé réalisable, appelé FT-MDnet, pour relier plusieurs types de réseaux de détection de piétons classiques avec notre proposition de réseau de réidentification ATLnet et MDnet, afin d'extraire des caractéristiques de réidentification à haute performance. FT-MDnet est bien adapté aux multiples tâches de suivi d'objets et sera utilisé comme réseau fondamental pour réaliser notre système de suivi.

— Dans la majorité des algorithmes de suivi de pointe, les informations à l'échelle spatiale des objets, comme leur position, leur taille, etc., sont les principaux indices pour l'association de données entre trames. Les caractéristiques de réidentification ou les descripteurs d'apparence sont considérés comme un moyen auxiliaire d'améliorer la robustesse de l'association dans les scènes encombrées où plusieurs objets peuvent être situés dans la même porte d'association à l'échelle spatiale. Lorsque les contraintes à l'échelle spatiale ne sont pas satisfaites, c'est-à-dire dans la scène de caméras rapides ou d'objets en mouvement rapide, les caractéristiques de réidentification perdent leur effet et les indices d'association à l'échelle spatiale sont invalidés.

Les caractéristiques de ré-identification sont sous-estimées en raison de leur absence de discrimination pour modéliser le problème de repérage de la piste perdue ou le problème de détection erronée. Les caractéristiques de réidentification pourraient être utilisées pour trouver les objets les plus semblables en apparence. Mais elle ne pourrait pas prendre de décisions d'association pour juger si les objets appartiennent ou non à la même identité, en particulier lorsqu'une cible a perdu sa détection ou son état de suivi.

Les dispositifs de réidentification formés sur des ensembles de données de caméras croisées pour s'adapter à diverses scènes et variations d'apparence contiennent généralement des redondances et perdent leur spécificité pour une scène de poursuite spécifique. Il existe principalement une contradiction entre l'universalité (usage général) et la spécificité (usage spécifique). Lors de la formation d'une fonction de ré-identification, l'objectif est d'obtenir des représentations de caractéristiques de haut rang pour toutes les scènes, ce processus fait naturellement perdre aux caractéristiques de ré-identification leur caractère discriminant. Lorsque nous appliquons des fonctions de réidentification pour l'association entre images à l'intérieur d'une caméra, nous cherchons à supprimer/répondre les redondances, afin de rendre les caractéristiques plus distinctes.

Cette thèse réévalue l'importance des caractéristiques de ré-identification et propose un convertisseur adaptatif en ligne pour convertir les caractéristiques de l'espace des caractéristiques de ré-identification en un espace de caractéristiques d'association. Les caractéristiques converties s'adaptant à une scène spécifique sont suffisantes pour prendre des décisions d'association qui pourraient améliorer de manière significative la continuité du suivi dans les scénarios mobiles.

— Les détecteurs d'objets sont généralement conçus et entraînés pour détecter des objets dans des images à une seule image. Sans utiliser les informations du contexte vidéo, il sera facile de provoquer des erreurs de détection, des fausses détections, du bruit de détection et d'autres problèmes de détection dans des scènes pratiques qui diffèrent des ensembles de données de formation.

Ces défauts de détection ne sont pas faciles à résoudre ou à corriger par le traqueur en ligne, car le mode de mouvement non linéaire des objets de vision est difficile à modéliser, à prévoir et à filtrer, et toute tentative de corriger les défauts de détection dégradera la robustesse des traqueurs. Le tracker en ligne reçoit passivement les résultats de la détection et dépend fortement de la justesse de la détection ; les défauts de détection affecteront directement la continuité et la précision de localisation des trajectoires.

Pour résoudre ce problème, nous proposons une solution appelée détection adaptative à la scène. L'idée est que les trajectoires du traqueur en ligne reliant la détection de plusieurs images passées et associant la détection de l'image actuelle pourraient confirmer la détection de l'image actuelle comme vraie positive avec un haut niveau de confiance. La transmission de ces informations au réseau de détection pourrait renforcer la confiance dans la détection d'objets spécifiques dans une scène spécifique. En appliquant cette solution, les détecteurs pourraient améliorer efficacement la détection des objets faibles afin d'améliorer la continuité du suivi tout en maintenant un faible taux de détection des faux négatifs.

Chapitre 2: Revue de la littérature : État de l'art

Objet Detectioin

La détection d'objets est une tâche de vision par ordinateur qui se développe rapidement. Elle constitue un point de départ pour la conception de plusieurs algorithmes de suivi d'objets et détermine presque la performance finale de suivi. À partir d'une image, sa tâche consiste à localiser les objets spécifiés par catégorie sur le plan de l'image et à donner simultanément des informations sur la catégorie. Les détecteurs à une étape les plus représentatifs actuellement sont les détecteurs de série YOLO [8, 9, 10]. YOLO a atteint un bon équilibre entre la vitesse d'inférence et la précision et a vraiment répondu aux exigences de la détection en temps réel. C'est le premier réseau de détection pris en compte lors de la conception d'une application du monde réel. Les détecteurs de la série RCNN [11, 12, 13, 14], réputés pour leur grande précision, sont les détecteurs à deux étapes les plus utilisés dans de nombreux domaines de recherche basés sur la vision, tels que la réidentification des personnes, la recherche de personnes et le suivi d'objets multiples. Les algorithmes YOLO et RCNN récemment publiés sont basés sur des ancres qui énumèrent de manière exhaustive toutes les boîtes englobantes et affinent les boîtes englobantes via un régresseur. Les méthodes basées sur des ancres nécessitent un processus de post-traitement, comme la non-suppression maximale (NMS), pour supprimer les boîtes englobantes qui se chevauchent ou qui sont en double. Récemment, un concept plus concis appelé "sans ancrage" a été proposé et commence à montrer ses avantages pour traiter le problème de la détection d'objets. Les détecteurs sans ancrage les plus typiques sont CornerNet [15] et CenterNet [16].

Réidentification et recherche de personnes

L'apprentissage global des traits apprend et extrait directement les représentations des traits sur l'image de la personne entière sans les contraintes de la partie. En raison de l'application originale des réseaux neuronaux profonds dans le problème de la classification des images, l'apprentissage global des caractéristiques est un choix primordial dès le plus jeune âge pour intégrer des techniques avancées d'apprentissage profond [32, 33]. Il est discriminant lorsque la détection de la personne permet de localiser avec précision le corps humain. Lorsque les images d'une personne souffrent d'un encombrement important du fond ou de fortes occlusions, l'apprentissage des caractéristiques au niveau partiel permet généralement d'obtenir de meilleures performances en exploitant les régions corporelles discriminantes [34, 35, 36]. Les solutions de AlignedReID [37], AlignedReID++ [38], sont

proposées pour résoudre le problème de l'image non alignée.Leur solution consiste à diviser une image en 8 bandes et à apprendre séparément la représentation des caractéristiques au niveau des parties pour faire correspondre dynamiquement les informations locales d'une image à l'autre. Dans [39], un modèle d'attention contrastive guidée par masque est proposé pour éliminer l'impact du fouillis de fond, formé avec une perte de triplet au niveau de la région. Réseau à granularité multiple (MGN) [40] propose d'utiliser une architecture de réseau à granularité multiple (CNN) profond pour les représentations d'entités locales et globales. La branche locale est constituée de segmentations à différentes échelles appelées granularité multiple pour représenter différentes régions du corps humain. En raison de l'avantage qu'elles présentent en termes de traitement des désalignements ou des occlusions, la plupart des méthodes de pointe développées récemment adoptent le paradigme de l'agrégation de caractéristiques qui combine les caractéristiques de niveau partiel et de niveau global.

suivi d'objets multiples

L'association de données en ligne est la partie centrale du suivi d'objets multiples, et elle est toujours conçue comme un problème d'affectation de données entre les pistes existantes et la détection de la prochaine arrivée. La résolution de l'association de données à l'échelle spatiale et dans l'espace des caractéristiques sont deux implémentations principales.

La plupart des algorithmes modernes de suivi d'objets multiples adoptent l'association de plusieurs indices pour résoudre le problème de l'association des données. Les descripteurs d'apparence ou les représentations d'apparence sont des indices robustes en raison de leurs contraintes non spatiales pour les scènes de mouvement de caméra. Lorsque l'association de boîtes limites engage des problèmes d'affectation, la fonction d'apparence est un moyen auxiliaire de prendre des décisions d'association. Après que l'apprentissage approfondi ait montré sa puissance dans la représentation des caractéristiques d'apparence, la représentation d'apparence basée sur l'apprentissage est souvent utilisée pour améliorer l'association de données. L'algorithme d'association de données en ligne le plus représentatif est **Deep-SORT** [62]. Un filtre de Kalman standard avec un mouvement à vitesse constante et un modèle d'observation linéaire propage l'identité d'une cible dans l'image suivante. Deep-SORT entraîne un CNN sur un ensemble de données de réidentification de personnes à grande échelle pour le rendre bien adapté à la description de l'apparence d'une personne. La distance de Mahalanobis, au lieu de l'IoU, entre les états de Kalman prévus et la détection nouvellement arrivée est utilisée pour incorporer la distance cosinusoïdale entre les voies et la détection afin de définir la matrice des coûts. Le solveur de la matrice des coûts est un algorithme hongrois.

Tracker++ [63] convertit un détecteur en un tracker en exploitant la régression bounding box du réseau de détection pour prédire la position d'un objet dans la prochaine trame d'arrivée. Tracker++ a étendu le réseau de détection avec deux extensions. **i**. un réseau siamois de ré-identification pour discriminer et ré-identifier les objets par leur apparence, en particulier pour le problème d'association de scènes encombrées. **ii**. appliquer une compensation de mouvement de caméra (CMC) pour les caméras en mouvement en alignant les images via l'enregistrement d'images en utilisant la maximisation du coefficient de corrélation amélioré [64].

Chapitre 3: Un cadre d'apprentissage de transfert surgelé pour la recherche de personnes

La mise en correspondance d'images de piétons recadrées manuellement entre les requêtes et les candidats, appelée ré-identification des personnes, a permis de réaliser des progrès significatifs avec les réseaux neuronaux convolutifs profonds. Récemment, un sujet appelé "recherche de personnes" a été proposé pour l'application de bout en bout des technologies de ré-identification. Il intègre la détection d'objets et la ré-identification de personnes et vise à la fois à localiser et à faire correspondre des piétons sur une galerie d'images brutes. Cependant, la conception et la mise en œuvre de ce type de réseau hybride sont difficiles et gourmandes en calculs dans des situations pratiques réelles. Afin d'affiner la conception et de faciliter la mise en œuvre, cet article propose un cadre d'apprentissage de transfert surgelé, appelé FT-MDnet, pour extraire les caractéristiques de ré-identification d'un réseau de détection préformé en deux étapes. Premièrement, en utilisant un mécanisme d'attention par canal, un réseau appelé réseau d'apprentissage de transfert adaptatif (ATLnet) est utilisé pour convertir les données de partage du réseau de détection sous-jacent en une carte de caractéristiques de ré-identification. Ensuite, un réseau de représentation des caractéristiques à branches multiples appelé réseau à descripteurs multiples (MDnet) est proposé pour extraire les caractéristiques de réidentification de la carte des caractéristiques de réidentification. La solution que nous proposons a été vérifiée sur différents types de réseaux de détection principaux, notamment YOLOv3, YOLOv4, Mask RCNN et CenterNet. Les résultats expérimentaux montrent que notre solution surpasse de loin toutes les autres solutions de recherche de personnes. Ils prouvent que les représentations des caractéristiques des réseaux de détection sont hautement compatibles avec la réidentification, et le cadre que nous proposons permet d'extraire efficacement ces caractéristiques. Pour encourager la poursuite des recherches, nous avons rendu notre cadre open source.

Proposition d'architecture de réseau mondial

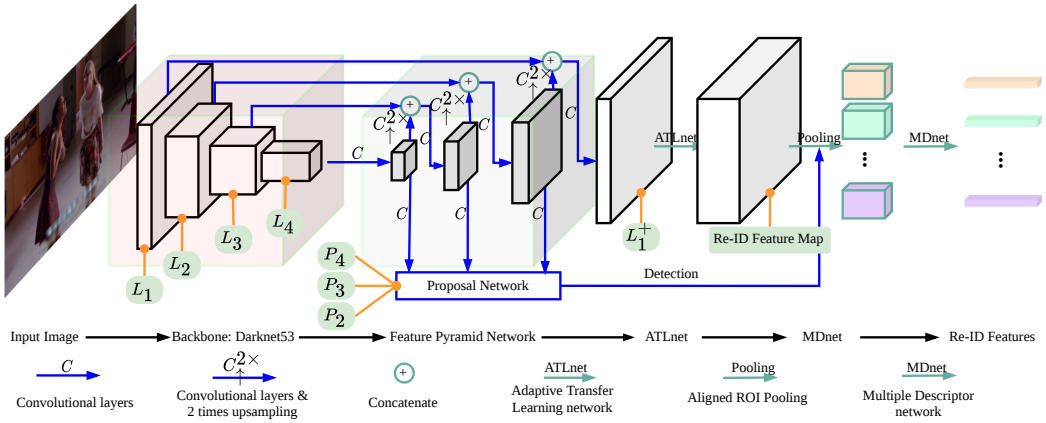


Figure 7.1 Architecture de réseau mondial.

Le réseau de détection moderne se compose généralement de trois parties, dont un réseau de base préformé sur ImageNet [19], un col (réseau de pyramides d'éléments) qui collecte les cartes d'éléments de plusieurs étages du réseau de base pour une représentation des éléments à plusieurs échelles, et un réseau de tête qui sert à la classification des objets et à la régression des boîtes englobantes ; voir Fig. 7.1.

Le réseau pyramidal de caractéristiques (FPN) [22, 92, 93] reliant le réseau fédérateur et le réseau de proposition est généralement utilisé pour un réseau de détection traitant le problème multi-échelle. Le FPN de YOLOv3 reçoit les cartes de caractéristiques du réseau fédérateur. De haut en bas ($L_4 \rightarrow L_2$), les cartes de caractéristiques sont échantillonnées et fusionnées vers le bas. Par exemple, la feature map L_4 a été traitée en premier lieu :

$$L_4^{'} = conv^{(5)}(L_4)$$
$$P_4 = conv^{(2)}(L_4^{'})$$

où $conv^{(i)}$ est un bloc de couches de convolution i. P_4 est une carte de proposition utilisée pour la détection d'objets. La carte des caractéristiques $L_4^{'}$ est suréchantillonnée et fusionnée avec L_3 par :

$$L_4^{''} = up^{2\times}(conv^{(1)}(L_4^{'}))$$
$$L_3^{+} = concat(L_4^{''}, L_3)$$

La proposition ATLnet

Nous adoptons un bloc de compression et d'excitation [95] pour rééquilibrer de manière adaptative les cartes de caractéristiques par canal. L'ATLnet pourrait s'écrire comme:

$$ATLnet(L_1^{+}) = conv_{dw}^{(1)}(L_1^{+} f_c^s(f_c^r(gap(L_1^{+}))))$$

où $conv_{dw}$ est une couche convolutionnelle séparable en profondeur. f_c est une couche entièrement connectée. *gap* est une couche de mise en commun de la moyenne globale.

La proposition MDnet

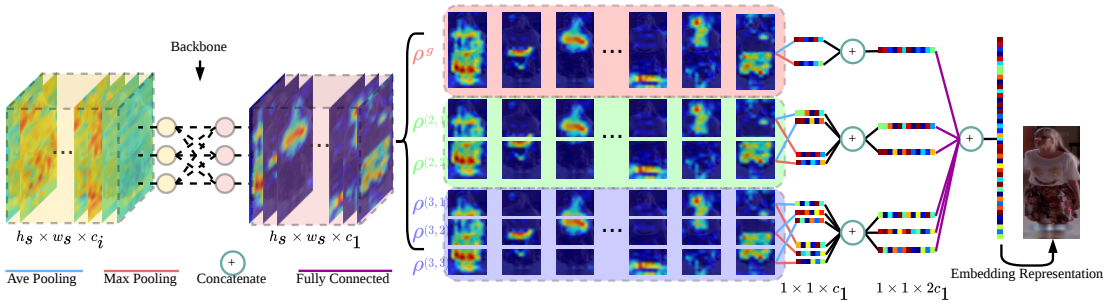


Figure 7.2 L'architecture de MDnet.

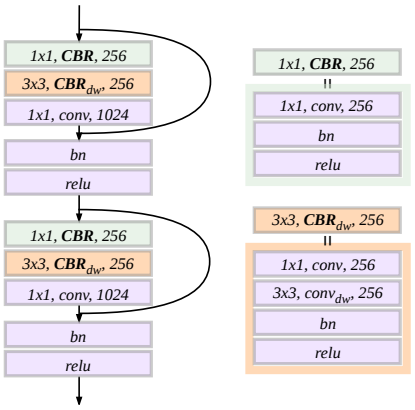


Figure 7.3 L'épine dorsale de MDnet.

L'épine dorsale de MDnet est constituée de deux unités résiduelles ; voir 7.3. Les couches convolutives avec un noyau à 3×3 sont remplacées par une couche convolutive séparable en profondeur. L'indice '$_{dw}$' signifie profondeur.

L'architecture de MDnet est présentée sur la figure 7.2. Les processus généraux sont les suivants :

— Le tenseur d'entrée est d'abord introduit dans l'épine dorsale pour obtenir la région de la caractéristique d'attention ϑ.

— En appliquant une architecture multi-branches, le tenseur ϑ est divisé en plusieurs parties. La branche locale comprend une branche à deux et une branche à trois branches qui divisent horizontalement le ϑ en sous-tenseurs de taille égale de 2 et 3.

— En incorporant la branche globale, les sous-tenseurs de 6 sont mis en commun par une couche de mise en commun maximale globale et une couche de mise en commun moyenne globale. Les vecteurs mis en commun sont concaténés par paires et alimentés séparément en couches entièrement connectées à 6 pour la réduction des dimensions.
— Concaténer les vecteurs pour obtenir la représentation finale de l'incrustation d'apparence.

Appariement et perte d'instance en ligne

La conception de la fonction de perte est inspirée de l'OIM [1]. En supposant qu'il y ait des piétons identifiés N, une structure de données appelée table de recherche (LUT): $\mathbf{L} \in \mathbb{R}^{N \times D}$ trié par id est utilisé pour garder toutes les identités, où D est la dimension des vecteurs de caractéristiques. Étant donné un vecteur de caractéristique $v_i \in \mathbb{R}^D$ où i est le id, il y a un et un seul échantillon positif dans la LUT, qui est désigné comme $\mathbf{L}_i$ (le i^{th} de $\mathbf{L}$). Les autres éléments de la liste sont tous négatifs en ce qui concerne v_i. La similitude en cosinus entre v_i et $\mathbf{L}_j$ est:

$$s(v_i, \mathbf{L}_j) = \frac{v_i^T \mathbf{L}_j}{\|v_i\| \, \|\mathbf{L}_j\|}$$

où $\|\cdot\|$ est la norme $l2$.

la fonction objective pourrait être écrite comme:

$$s^{obj}(v_i, \mathbf{L}_j) = \begin{cases} 1 & \text{if } j = i \\ -1 & \text{if } j \neq i \end{cases}$$

Étant donné les cibles étiquetées M $\mathbf{V} \in \mathbb{R}^{M \times D}$, la similarité est : $S_{ij} = s(\mathbf{V}_i, \mathbf{L}_j)$, où S_{ij} est l'élément de ligne i^{th} et de colonne j^{th} de la matrice de similarité $S \in \mathbb{R}^{M \times N}$. La matrice d'objectif correspondante est $S^{obj} \in \mathbb{R}^{M \times N}$. Nous effectuons tout d'abord un softmax sur chaque ligne de la matrice de similarité:

$$\tilde{S}_i = \frac{exp(S_i t)}{\sum_{j=1}^{N} exp(S_{ij} t)}$$

où S_i signifie la ligne i^{th} de S. La nouvelle matrice d'objectifs est:

$$\tilde{S}_{ij}^{obj} = \begin{cases} 1 & \text{if } S_{ij}^{obj} = 1 \\ 0 & \text{if } S_{ij}^{obj} = -1 \end{cases}$$

La perte du réseau pourrait s'écrire comme suit:

$$loss = -\frac{1}{M} \sum_{i=1,j=1}^{M,N} log(\tilde{S}_{ij} \tilde{S}_{ij}^{obj})$$

À la fin de chaque itération de lot, mettez à jour la LUT par:

$$\mathbf{L}_n \leftarrow \gamma \mathbf{L}_n + (1-\gamma) v_n$$

où $\gamma \in [0, 1]$.

Expériences et évaluation

Détails de la mise en œuvre

Notre solution est mise en œuvre sur Keras avec le backend TensorFlow. Pour la formation au réseau, nous utilisons une carte graphique Geforce GTX 1080Ti avec une mémoire de 11G. Nous ajoutons des régularisateurs de $l2$ aux paramètres de la convolution et des couches entièrement connectées avec une valeur de $0,0005$. Le ratio de réduction τ dans ATLnet est fixé à 16. Le modèle d'échantillonnage de la couche de mise en commun du retour sur investissement alignée est fixé à 24×12. Le taux de mise à jour de γ de LUT est fixé à $0,5$. Nous utilisons une taille de lot de 6. Nous définissons une époque comme une traversée de l'ensemble de l'entraînement. Le taux d'apprentissage est fixé à $0,01$ dans les premières époques à 20, puis le taux d'apprentissage diminue à $0,001$ pour les époques à 20 restantes. Pendant la formation, nous gelons l'ensemble du réseau de détection.

Résultats de l'évaluation

Table 7.1 Résultats de la comparaison d'évaluation pour YOLOv3 sur l'ensemble de données CUHK-SYSU

Methods	Years	CUHK-SYSU (50)		CUHK-SYSU (100)		Input Size
		mAP (%)	top-1 (%)	mAP (%)	top-1 (%)	
OIM[1]	2017	81.9	82.6	78.0	78.7	600×600
LCGPS[53]	2019	87.8	-	84.1	86.5	720×576
QEEPS[54]	2019	-	-	88.9	89.1	900×900
MGTS[48]	2018	84.8	-	83.0	83.7	-
CLSA[46]	2018	-	-	87.2	88.5	-
DHFF[47]	2019	-	-	90.2	91.7	-
YOLOv3-MDnet	-	90.5	91.8	89.2	90.6	608×608
YOLOv3-MDnet^{+}	-	**94.1**	**95.3**	**93.2**	**94.5**	608×608
YOLOv3-MDnet$^{+}_{+}$	-	92.2	92.9	91.4	92.0	608×608

Sur la base de ces expériences, nous avons pu constater les particularités suivantes de la solution proposée:

Table 7.2 Résultats de la comparaison d'évaluation pour YOLOv3 sur l'ensemble de données PRW

Methods	Years	PRW		Input Size
		mAP(%)	top-1(%)	
OIM[1]	2017	36.9	75.7	900 × 900
LCGPS[53]	2019	33.4	73.6	720 × 576
QEEPS[54]	2019	37.1	76.6	900 × 900
MGTS[48]	2018	32.6	72.1	-
CLSA[46]	2018	38.7	65.0	-
DHFF[47]	2019	41.1	70.1	-
YOLOv3-MDnet	-	43.4	83.0	608 × 608
YOLOv3-MDnet^{+}	-	52.0	86.0	608 × 608
YOLOv3-MDnet$^{+}_{+}$	-	**54.9**	**87.2**	608 × 608

Table 7.3 Résultats de la comparaison des évaluations pour d'autres détecteurs classiques sur CUHK-SYSU

Methods	CUHK-SYSU (50)		CUHK-SYSU (100)		Input Size
	mAP (%)	top-1 (%)	mAP (%)	top-1 (%)	
YOLOv4-MDnet	91.6	92.1	89.7	90.2	608 × 608
MRCNN-MDnet	92.7	93.8	91.6	92.7	1024 × 1024
YOLOv4-MDnet^{+}	**94.6**	**95.2**	**93.7**	**94.5**	608 × 608
CenterNet-MDnet^{+}	93.9	95.0	92.9	94.0	608 × 1088

Table 7.4 Résultats de la comparaison d'évaluation pour d'autres détecteurs principaux sur l'ensemble de données PRW

Methods	PRW		Input Size
	mAP(%)	top-1(%)	
YOLOv4-MDnet	41.2	80.7	608 × 608
MRCNN-MDnet	47.4	83.3	1024 × 1024
YOLOv4-MDnet^{+}	**55.1**	**86.3**	608 × 608
CenterNet-MDnet^{+}	54.6	86.3	608 × 1088

— Le cadre que nous proposons fonctionne efficacement avec la dorsale YOLOv3
— Meilleur est le réseau de détection, meilleure est la précision de la recherche de la personne
— La formation de bout en bout YOLOv3-MDnet^{+} est sujette à la surcharge

Expériences prolongées

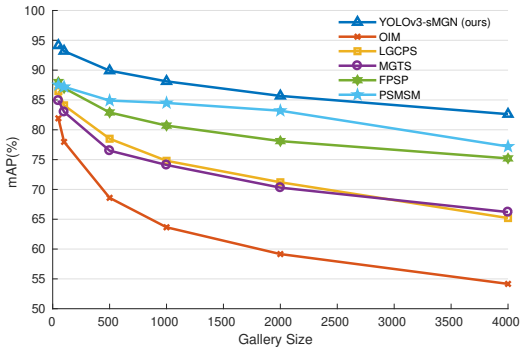


Figure 7.4 L'impact de la taille de la galerie sur la performance

Pour évaluer pleinement les performances de CUHK-SYSU, la Fig. 7.4 montre la précision de la recherche avec différentes tailles de galeries–[50, 100, 500, 1000, 2000, 4000]. Nous recueillons les résultats d'évaluation de l'OIM, du LGCPS, du FPSP [45], de la CLSA [46], de la DHFF [47] et de la MGTS [48] à partir des documents correspondants ; parmi ceux-ci, le FPSP, la CLSA, la DHFF, la MGTS sont des solutions IPS. On peut constater que, dans l'ensemble des méthodes énumérées, la précision diminue rapidement à mesure que la taille de la galerie augmente. La solution que nous proposons, YOLOv3-MDnet^{+}, surpasse de loin toutes les autres solutions IPS et JPS sur toutes les tailles de galerie.

Nous avons visualisé les résultats de la recherche en utilisant t-SNE [96], voir Fig. 7.5. Nous avons projeté toutes les caractéristiques Re-ID des piétons identifiables de PRW sur un plan 2D. Il y a 450 de piétons, soit 16344 d'images de piétons. Chaque image de piéton correspond à un point ; les points de même couleur indiquent le même piéton. On peut voir que notre solution a un assez bon effet de regroupement. Nous avons sélectionné une région à 6 sur la carte des nuages de points et nous avons tiré les trois conclusions suivantes:

— Les régions 1 et 2 sont bien distinctes sur la carte des nuages de points. Les piétons des régions 1, 2 portant des vêtements de style similaire sont bien distincts par la couleur. Cela signifie que le réseau de détection préformé a retenu les informations de couleur.
— Dans la région 3, 4, deux piétons d'apparence similaire se distinguent par les dé-

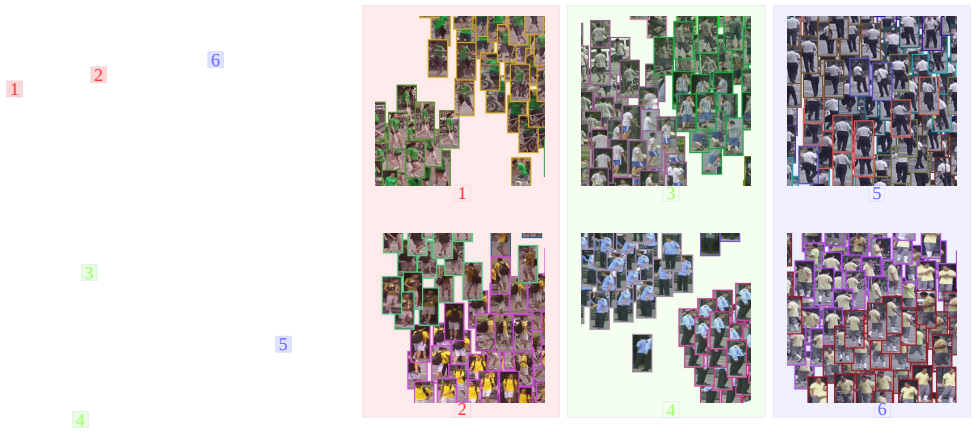


Figure 7.5 Visualisation du t-SNE sur le jeu de données PRW.

tails. Dans la région 3, l'objet tenu à la main et les chaussures fournissent des informations discriminantes. Dans la région 4, les agents de sécurité se distinguent par des uniformes à manches longues et à manches courtes.

— Dans les régions 5 et 6, les piétons d'apparence similaire sont bien regroupés. Mais ils sont susceptibles de donner des résultats de recherche erronés.

Conclusions

La recherche de personnes intégrant la détection d'objets et la réidentification de personnes est une tâche très difficile en raison des difficultés de conception du réseau de détection. Ce chapitre a proposé un cadre d'apprentissage de transfert surgelé qui forme un réseau de ré-identification sous-jacent à un détecteur préformé tout en gardant l'ensemble du réseau de détection gelé. Le réseau de ré-identification composé d'ATLnet et de MDnet a été proposé pour apprendre la représentation des caractéristiques d'apparence à plusieurs échelles et en plusieurs parties. Le cadre proposé, mis en œuvre sur plusieurs détecteurs courants, dont YOLOv3, YOLOv4, Mask RCNN et CenterNet, a permis d'améliorer remarquablement les performances des ensembles de données de recherche de deux personnes CUHK-SYSU et PRW. Les résultats expérimentaux démontrent que les réseaux de détection préformés répertoriés conservent suffisamment d'informations discriminantes pour une nouvelle identification et les réseaux ATLnet et MDnet proposés, de conception peu profonde, pourraient efficacement extraire des caractéristiques discriminantes du réseau de détection gelé. Le cadre que nous proposons, qui est facile à intégrer les réseaux de détection de pointe, pourrait grandement simplifier la conception des solutions de recherche conjointe de personnes. Le code source de notre travail (comprenant la formation, l'évaluation, la visualisation et une simple application de recherche de personnes)

est disponible sur github. [1]

Chapitre 4: Association de données adaptatives de scène

L'association des données dans la plupart des algorithmes de suivi d'objets multiples existants est modélisée comme un problème d'association entre trames. Il existe principalement deux types de contraintes utilisées pour faire correspondre la détection inter-trames. Ils sont:

— **Contraintes à l'échelle spatiale**, la distance entre les cases délimitant les cadres qui contribuent aux contraintes d'association à l'échelle spatiale.
— **Contraintes d'espace**, les caractéristiques de réidentification contribuant à la contrainte de similarité d'apparence dans l'espace des caractéristiques.

Le chapitre 7 a proposé une architecture de recherche de personnes de bout en bout qui intègre la détection d'objets et l'extraction de caractéristiques de réidentification. Il fournit un modèle de base pour la conception du système de suivi d'objets multiples.

Dans de nombreux algorithmes de pointe de suivi d'objets multiples [68, 62, 63], les caractéristiques de réidentification sont considérées comme des moyens supplémentaires pour améliorer la robustesse du système de suivi et ne sont presque jamais utilisées indépendamment des indices à l'échelle spatiale pour prendre des décisions d'association. MAT[57] fait valoir qu'en raison des détections partielles bruyantes et de l'absence de contraintes spatio-temporelles, la fonction de réidentification n'est pas fiable, prend du temps et ne peut toujours pas traiter les faux négatifs des objets occultés et flous. Les raisons de cette sous-évaluation sont les suivantes:

— En cas de variation importante de la posture ou de l'apparence, ces fonctions de ré-identification inter-caméras sont susceptibles de perdre leur fonction. Cela entraîne l'application de la ré-identification dans de nombreux algorithmes de suivi, qui prend du temps, n'est pas fiable et est limitée dans certaines scènes de suivi "faciles".
— Les dispositifs de réidentification bien conçus sont conçus pour s'adapter à diverses scènes et variations d'apparence ; ils contiennent techniquement des redondances et manquent de signification. La distinction entre la similarité des vraies paires de concordances et la similarité des fausses paires de concordances est insignifiante, donc peu fiable pour prendre des décisions d'association.

Ce chapitre s'engage à discuter et à résoudre le deuxième problème—**comment utiliser les caractéristiques de réidentification indépendamment des indices à l'échelle**

1. https://github.com/RuaHU/TLfPS

spatiale pour prendre des décisions d'association.

L'algorithme d'association

Algorithm 5: Association()

Input : D_t, F_t, T_{t-1}

Output: $M_{t-1|t}$

`/* obtenir la matrice de similarité d'apparence et la matrice de distance */`

foreach $track_{t-1}^i \in T_{t-1}$, $det_t^j \in D_t$ **do**

$W_{t-1|t}^{reid} = \{w_{i,j}^{reid}\}_{i=1,j=1}^{n,m}$;

$w_{i,j}^{reid} = \frac{\sum f_{track_{t-1}^i} f_{det_t^j}}{\left\| f_{track_{t-1}^i} \right\|_2 \left\| f_{det_t^j} \right\|_2}$;

$W_{t-1|t}^{dist} = \{w_{i,j}^{dist}\}_{i=1,j=1}^{n,m}$;

$w_{i,j}^{dist} = 1 - \frac{2}{1+\exp(-d_{i,j})}$;

KalmanFilter : $track_{t-1}^i \mapsto track_{t-1|t}^i$;

$d_{i,j} = (det_t^j - b_{track_{t-1|t}^i})^T (H^T P_{t-1|t} H)^{-1} (det_t^j - b_{track_{t-1|t}^i})$;

end

`/* fusionner plusieurs indices */`

$W_{t-1|t} = \alpha W_{t-1|t}^{reid} + (1-\alpha) W_{t-1|t}^{dist}$;

`/* résoudre la matrice de poids en utilisant le HA */`

HA : $W_{t-1|t} \mapsto M_{t-1|t}$

où D_t sont les détections dans la trame t. F_t sont les vecteurs de caractéristiques de réidentification correspondants. T_{t-1} sont les pistes dans la trame $t-1$. $M_{t-1|t}$ sont les correspondances d'association entre D_t et T_{t-1}. $W_{t-1|t}^{reid}$ est la matrice de poids d'association en utilisant des vecteurs de caractéristiques de réidentification. $W_{t-1|t}^{dist}$ est la matrice de poids d'association en utilisant des boîtes englobantes à l'échelle spatiale.

Vérifier la solidité de l'association des caractéristiques de réidentification

Nous utilisons les séquences de formation étiquetées de MOT17 pour l'évaluation.

La précision de l'association de la fonction de réidentification dépasse 97% sur toutes les séquences vidéo. La précision de la distance de jointure et de la ré-identification dépasse 98% sur toutes les vidéos. Lorsque l'intervalle entre les images augmente, l'association de

la fonction de ré-identification montre ses avantages dans une association robuste pour les vidéos à faible fréquence d'images. Intuitivement et expérimentalement, l'application de contraintes à l'échelle spatiale pour l'association nécessite une forte hypothèse que le mouvement inter-trame de l'objet est faible. Une fois que l'intervalle entre les images augmente ou que la caméra change de position ou de direction, la précision de l'association se dégrade rapidement, ce qui entraîne un manque de robustesse. L'association spatiale des objets est plus robuste pour toutes les fréquences d'images et toutes les scènes.

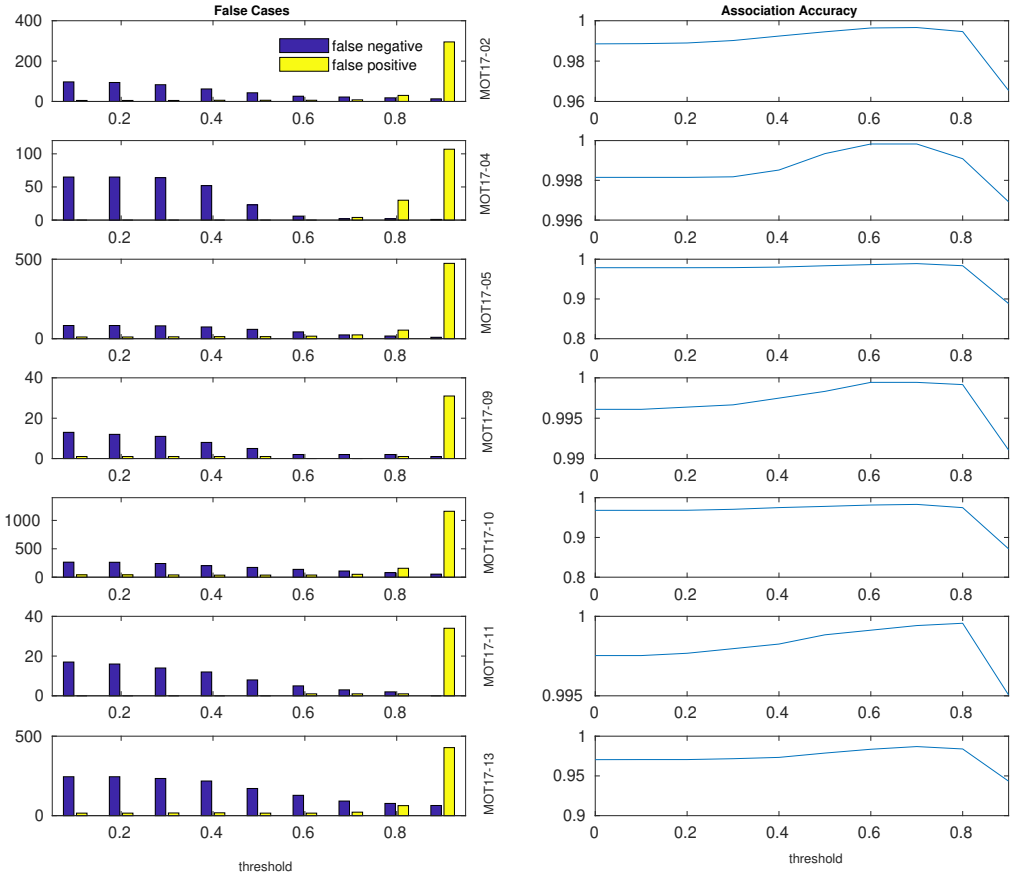


Figure 7.6 Caractéristique de réidentification Précision d'association avec un seuil différent.

Nous changeons le seuil de $0 \rightarrow 1$ avec un intervalle de $0,1$ et fixons l'intervalle de trame à 1. Les résultats sont présentés dans la figure 7.6. Nous traçons le nombre de cas de faux positifs et de faux négatifs. On peut voir sur la figure que lorsque le seuil augmente, le nombre de cas faux-positifs augmente et le nombre de cas faux-négatifs diminue. Lorsque le seuil est de $0,7$, le nombre total de cas de faux positifs est minime. Dans ce cas, la précision de l'association inter-trames de l'utilisation des caractéristiques

de ré-identification dépasse 98%.

Nos expériences prouvent que les caractéristiques de réidentification peuvent fournir une grande précision d'association. Elle présente un avantage incomparable pour comparer l'association de boîtes englobantes à l'échelle spatiale dans les vidéos à faible fréquence d'images. Cependant, le problème existant est que lorsqu'une cible a perdu sa détection dans l'image suivante. La fonction de ré-identification ne peut pas donner une discrimination suffisante pour juger si la cible est perdue ou non.

Améliorer la discrimination de la caractéristique de réidentification et résoudre le problème d'association dans l'espace de caractéristiques est un problème à résoudre dans ce chapitre.

Proposition d'une association de données adaptées à la scène

Nous définissons un convertisseur qui convertit les vecteurs de caractéristiques de ré-identification de l'espace de caractéristiques de ré-identification en espace de caractéristiques d'association:

$$\boldsymbol{\chi}^{asso} = \mathbf{W}^{reid|asso}\boldsymbol{\chi}^{reid} + \mathbf{B}$$

où $\mathbf{W}^{reid|asso} \in \mathbb{R}^{n,m}$ est la matrice de conversion. $\boldsymbol{\chi}^{asso}$ est un espace de caractéristiques d'association qui a m *bases*. $\mathbf{B}$ est un vecteur de biais dimensionnel m.

L'architecture globale de l'association scène-adaptation est présentée dans 7.7. Nous définissons d'abord la matrice d'association (matrice de poids en utilisant des caractéristiques converties) comme:

$$W_{t-1|t} = \{w_{i,j}\}_{i=1,j=1}^{n,m}$$
$$w_{i,j} = \frac{\sum f_{track^i}^{trans} f_{det^j}^{trans}}{\left\|f_{track^i}^{trans}\right\|_2 \left\|f_{det^j}^{trans}\right\|_2}$$

où:

$$f_{track^i}^{trans} = \mathbf{T}_{t-1|t}(f_{track^i}) = f_{track^i}\mathbf{W}_{t-1|t}^{reid|asso} + \mathbf{B}_{t-1|t}$$
$$f_{det^j}^{trans} = \mathbf{T}_{t-1|t}(f_{det^j}) = f_{det^j}\mathbf{W}_{t-1|t}^{reid|asso} + \mathbf{B}_{t-1|t}$$

nous ajoutons une colonne et une ligne supplémentaires de valeurs scalaires pour élargir la matrice de poids:

$$\bar{W}_{t-1|t} = \begin{pmatrix} & & & \alpha \\ & W_{t-1|t} & & \vdots \\ & & & \alpha \\ \alpha & \dots & \alpha & \alpha \end{pmatrix}$$

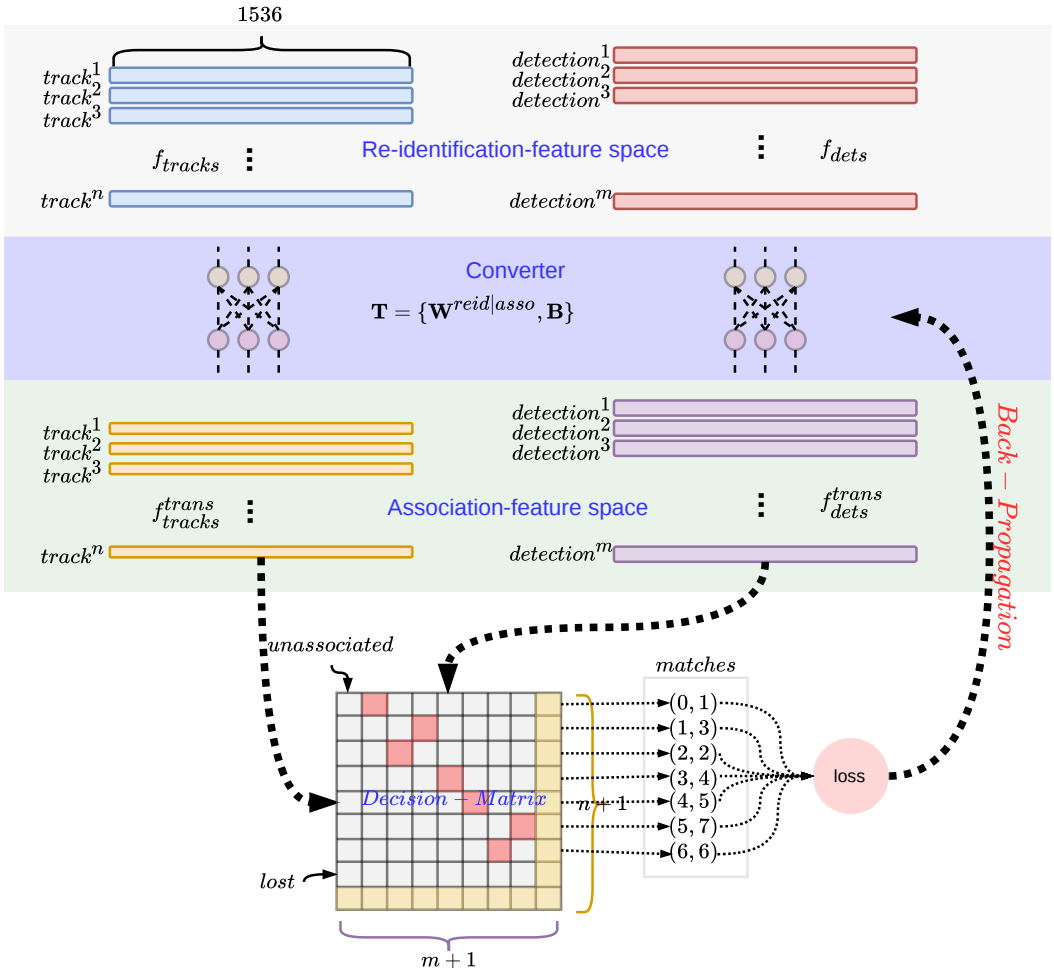


Figure 7.7 L'Association pour l'architecture des données adaptées à la scène.

Nous effectuons des softmax pour chaque ligne et chaque colonne de la matrice $\bar{W}_{t-1|t}$. Les matrices softmax sont désignées par:

$$\tilde{W}^{row}_{t-1|t} = softmax(\bar{W}_{t-1|t}\tau)^{row}$$
$$\tilde{W}^{col}_{t-1|t} = softmax(\bar{W}_{t-1|t}\tau)^{col}$$

Perte et rétropropagation

L'objectif de la fonction de perte est de maximiser le poids des points appariés et de minimiser le poids des points non appariés des deux matrices de décision:

$$w_{i,j} = \begin{cases} 1 & if\ w_{i,j}\ matched \\ 0 & otherwise \end{cases}$$

Supposons qu'il y ait N $maths = \{(w_i^{row}, w_i^{col})\}_{i=1}^{N}$. La fonction de perte est :

$$loss = -\frac{1}{2N}\sum_{i=1}^{N}(\log w_i^{row} + \log w_i^{col})$$

L'ensemble du processus est différenciable, les équations de la rétropropagation s'écrivent comme:

$$\mathbf{W}_{t|t+1}^{reid|asso} = \mathbf{W}_{t-1|t}^{reid|asso} + \eta\frac{\delta_{loss}}{\delta_{\mathbf{W}_{t-1|t}^{reid|asso}}}$$

$$\mathbf{B}_{t|t+1} = \mathbf{B}_{t-1|t} + \eta\frac{\delta_{loss}}{\delta_{\mathbf{B}_{t-1|t}}}$$

Visualiser les résultats de l'association d'une scène en mouvement rapide

La figure 7.8 montre les résultats de l'association en utilisant la méthode que nous proposons; les cases délimitées par la même couleur indiquent la même personne. Les matrices détaillées des poids d'association sont indiquées. On peut voir que l'AFWM présente une grande similitude discriminante avec le RFWM. Les deux matrices de décision illustrent des associations claires qui pourraient être utilisées pour prendre des décisions d'association.

Conclusions

Ce chapitre a examiné l'application de la fonction de réidentification pour l'association de données. Contrairement à d'autres méthodes de suivi de pointe qui considéraient la fonction de réidentification comme un moyen complémentaire d'améliorer la précision de l'association à l'échelle spatiale, nous réévaluons l'importance de la fonction de réidentification sur la base de notre précédent projet de réseau de recherche de personnes. Nous menons des expériences pour prouver que les fonctions de ré-identification sont plus robustes que l'association de boîtes englobantes à l'échelle spatiale dans toutes les scènes de suivi et toutes les vidéos à fréquence d'images. Nous avons découvert que le problème essentiel qui entrave l'application de la fonction de ré-identification est son manque d'importance dans le problème de l'association. Nous proposons un convertisseur adaptatif de scènes d'apprentissage en ligne qui convertit en ligne les caractéristiques de ré-identification en caractéristiques d'association. Nos expériences ont prouvé que les caractéristiques d'association pouvaient fournir suffisamment de similitudes significatives entre les correspondances positives et négatives ; il est capable de prendre des décisions d'association fiables indépendamment des indices à l'échelle spatiale.

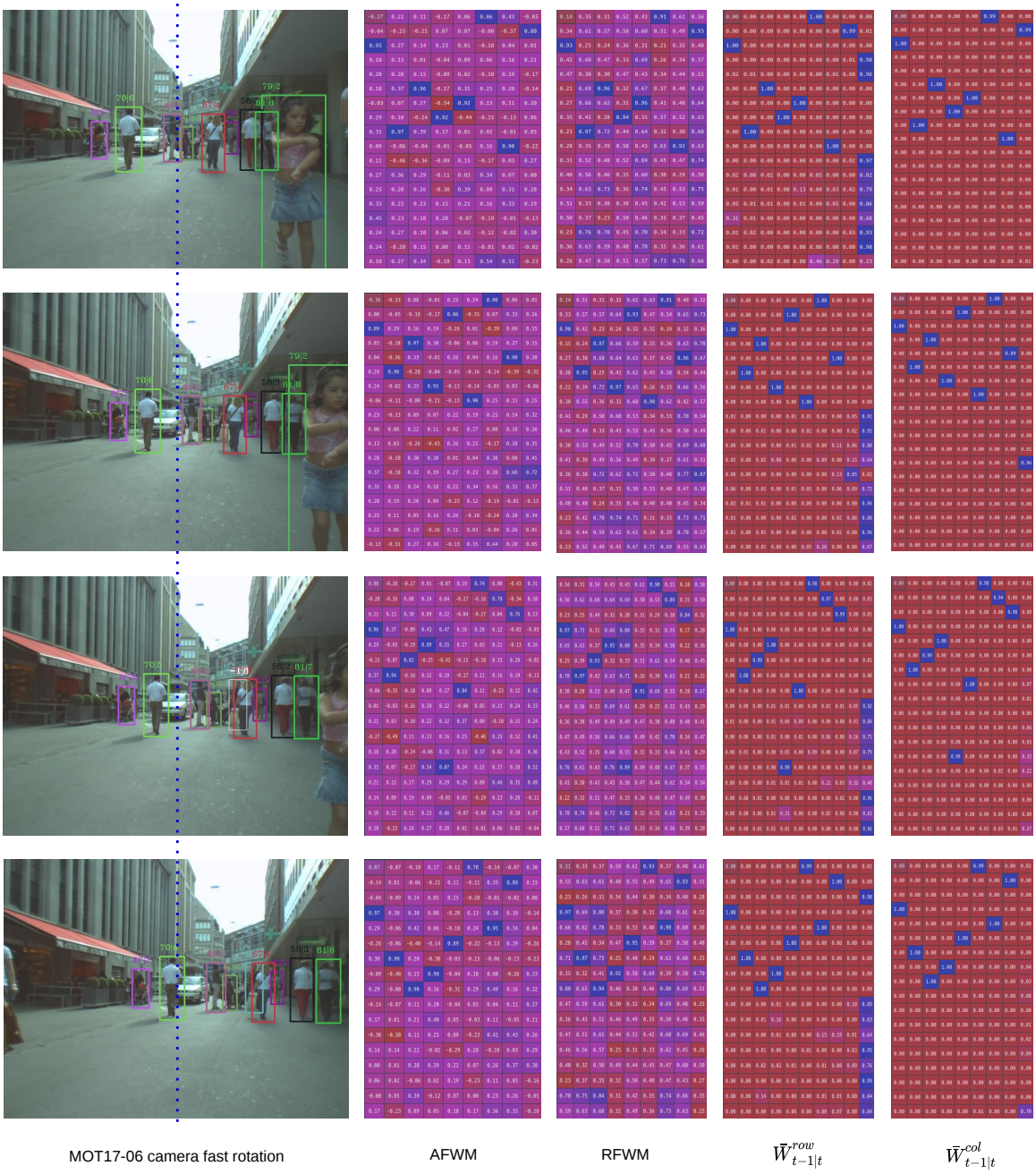


Figure 7.8 Illustrer les résultats de l'association de scènes en mouvement rapide.

Chapitre 5: la détection adaptative de scènes et l'architecture globale du suivi d'objets multiples

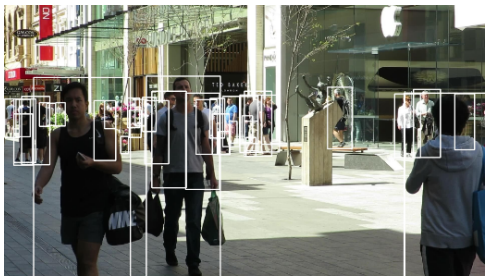
Compte tenu du problème de la mauvaise détection, l'idée de notre proposition de détection adaptée à la scène est de renvoyer les résultats du suivi au réseau de détection afin d'améliorer la confiance dans la détection d'objets spécifiques.

Dans le paradigme traditionnel du suivi par détection, la détection et le suivi des objets sont considérés comme deux modules indépendants. Le module de suivi/association reçoit passivement les résultats de la détection sans retour d'information. Des erreurs de détection imprévisibles peuvent se produire dans différentes scènes d'application, telles que la mauvaise détection d'un objet et le bruit de localisation d'un objet. Ces défauts de détection sont difficiles à traiter par le tracker. D'autre part, les détecteurs effectuent une détection par image ; la qualité des résultats de détection ne dépend que de l'image en cours et ne tient pas compte des contraintes du contexte vidéo. Les algorithmes de poursuite d'objets multiples relient la trajectoire spatio-temporelle des cibles dans une séquence d'images continue ; le filtre de poursuite contraint la position spatiale des objets pour affiner la précision de la localisation et pourrait confirmer l'existence d'objets réellement positifs. Une idée naturelle est de savoir si nous pouvons utiliser les résultats du suivi pour optimiser le réseau de détection.

Cette thèse choisit CenterNet comme réseau de détection sous-jacent; l'implémentation du tracker est basée sur notre précédent réseau de recherche de personnes proposé—CenterNet-MDnet.

Nous avons l'intention de transmettre les résultats du suivi au réseau de détection afin d'améliorer l'adaptabilité du détecteur à la scène actuelle. CenterNet dispose de trois sorties, dont une carte thermique $\hat{I}^{hm}$ pour prédire le centre des objets, une carte de régression de hauteur et de largeur $\hat{I}^{wh}$ pour prédire la hauteur et le poids des objets, et une carte de régression de position $\hat{I}^{reg}$ pour améliorer la précision de la localisation. Comme nous l'avons vu plus haut, en raison de la non-linéarité du mouvement des objets, le filtre de Kalman ne peut pas réduire le bruit de détection et peut introduire un bruit non linéaire, les résultats du filtrage du tracker ne sont pas adaptés pour optimiser la hauteur, la largeur et la position de la détection. Au contraire, le fait de renvoyer la carte thermique au réseau de détection et de mettre à jour le réseau de détection ne détruit pas essentiellement les performances du détecteur si nous trouvons suffisamment d'échantillons positifs et d'échantillons négatifs.

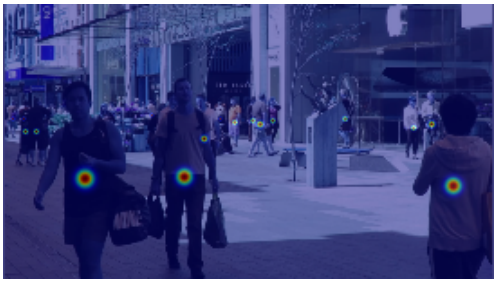
Échantillons positifs et échantillons négatifs



(a) detection and prediction

(b) red: negative samples; blue: invalid samples

(c) red: negative samples (except the center point)



(d): gaussian kernel

Figure 7.9 Illustration des échantillons positifs et des échantillons négatifs.

La figure 7.9 montre les échantillons d'entraînement.

- Image *(a)* : les cases blanches délimitées sont les détections générées par le réseau de détection et les prédictions générées par le tracker.
- Image *(b)* : les régions rouges hors de la couverture de toutes les cases limites sont confirmées comme étant des échantillons négatifs. Les régions bleues sont des échantillons non confirmés (non valides) et ne seront pas utilisées comme échantillons de formation.
- Image *(c)* : les cases de délimitation de la détection confirmées par le tracker comme étant une détection réelle marqueront la région correspondante comme négative (sauf le point central, qui est un échantillon positif). Les échantillons négatifs seront utilisés pour transformer la région non valide de l'image *(b)*. (échantillons invalides transformés en échantillon négatif et en échantillon positif).
- Image *(d)* : les détections confirmées sont représentées par des noyaux gaussiens.

Nous désignons la carte thermique par $I^{hm}_{x,y}$. La surface effective de la carte thermique pourrait alors s'écrire comme suit:

$$I^{e}_{x,y} = \begin{cases} 1 & if\ I_{x,y} > 0\ and\ I_{x,y} valid \\ 0 & otherwise \end{cases}$$

où x, y sont les indices des éléments. Les pixels en dehors des régions non valides sont des pixels valides.

La conception de la fonction de perte

Dans la mise en œuvre originale de CenterNet, tous les pixels de la carte thermique sont des échantillons négatifs, à l'exception du point central du noyau gaussien. Dans la méthode que nous proposons, seules les zones effectives de la carte thermique sont prises en considération. La fonction de perte de rétroaction est:

$$loss^{hm} = -\frac{1}{N}\sum_{x,y}\begin{cases}(1-(\hat{I}^{hm}_{x,y})^{\alpha})log(\hat{I}^{hm}_{x,y}) & if\ I^{hm}_{x,y}=1\\ (1-I^{hm^{\beta}}_{x,y})(\hat{I}^{hm}_{x,y})^{\alpha}log(1-\hat{I}^{hm}_{x,y})I^{e}_{x,y} & otherwise\end{cases}$$

où $\hat{I}^{hm}_{x,y}$ est la carte de chaleur prédite générée par le détecteur. où $I^{hm}_{x,y}$ est la carte de chaleur de vérité du sol générée par le tracker. Et $I^{e}_{x,y}$ est la surface effective de la carte thermique.

L'archétecture globale du système de suivi adapté à la scène

Nous avons divisé CenterNet en deux parties—$CenterNet^{-}$ et $conv^{hm} = \{\mathbf{K}^{hm}, \mathbf{B}^{hm}\}$. $conv^{hm}$ est la dernière couche convolutionnelle qui produit $\hat{I}^{hm}_{x,y}$. K^{hm} et B^{hm} sont le noyau convolutif et le vecteur de biais correspondants. La prédiction de la carte thermique pourrait s'écrire:

$$\hat{I}^{hm}_{x,y} = conv^{hm}(CenterNet^{-}(I)) \tag{7.1}$$

où I est l'image d'entrée. L'architecture globale de notre système de suivi adaptatif des scènes proposé est illustrée à la figure 7.10. Les processus globaux sont énumérés comme suit:

1 **Pretraitement** : L'entrée du système de suivi est constituée d'images ; les images sont d'abord redimensionnées à une forme spécifique pour s'adapter à la forme d'entrée du réseau de détection.

2 **Détection d'objet** : L'image redimensionnée calculée par le réseau de détection pour générer des boîtes de délimitation de la détection. Au fait, la carte thermique $\hat{I}^{hm}_{x,y}$ prédite par la couche convolutionnelle $conv^{hm}$ est également produite.

3 **extraction des caractéristiques de réidentification**: Le sous-réseau de réidentification $ATLnet$ et $MDnet$ extrait les caractéristiques d'apparence pour chaque boîte englobante détectée.

4 **Convertisseur adapté à la scène**: Le convertisseur a été effectué pour convertir les caractéristiques de réidentification de la détection et des pistes en caractéristiques d'association.

5 **Association Cascade**: L'association consiste en deux étapes en cascade. La première étape consiste à obtenir des $matches^{reid}$ en résolvant les deux $Decision-Matrix$ 4.24. La deuxième étape consiste à obtenir $matches^{join}$ en résolvant la

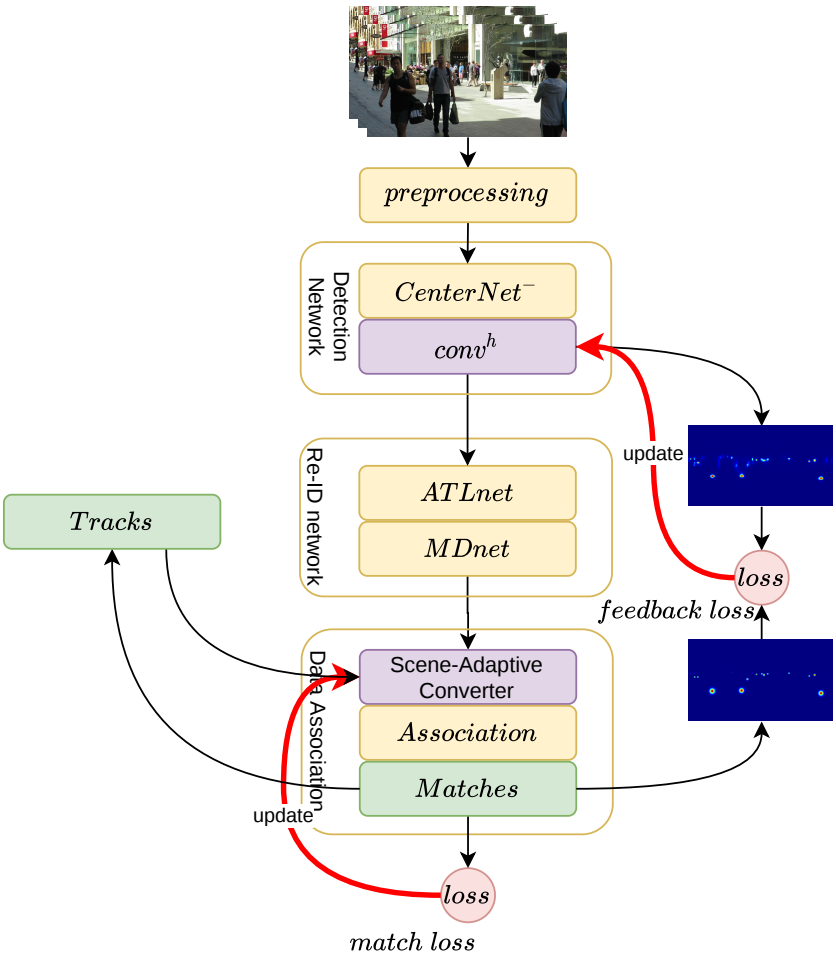


Figure 7.10 L'archétecture globale du système de suivi adapté à la scène.

matrice de poids d'association de l'échelle spatiale de jointure et de l'espace de caractéristiques 4.17.

6 **Mise à jour des pistes**: Utilisation des résultats de la comparaison pour mettre à jour la liste des pistes.

7 **Mise à jour du convertisseur**: En fonction des correspondances, la perte de correspondance 4.26 back se propage au convertisseur adapté à la scène 4.27 pour mettre à jour les poids.

7 **Créer une carte thermique**: Les pistes confirmées comme étant réellement positives seront utilisées pour créer une carte thermique gaussienne du noyau $I^{hm}_{x,y}$.

8 **Mise à jour du réseau de détection**: Obtenir la perte de rétroaction 5.3 et la propager en retour au réseau de détection pour mettre à jour les poids de $conv^{hm}$

Table 7.5 Résultats généraux de l'évaluation

Tracker	**MOTA↑**	**IDF1↑**	**MT↑**	**ML↓**	**ID SW↓**	**Frag↓**
FairMOT[68]	73.7	72.3	1017	408	3303	8073
MAT[57]	69.5	63.1	1032	444	2844	**3726**
CTTracker[101]	67.8	64.7	816	579	3039	6102
CSTrack[102]	**74.9**	**72.6**	978	411	3567	7668
SATracker(Ours)	73.12	72.39	1023	375	1569	4716
SATracker+(Ours)	74.25	72.25	**1095**	**294**	**1416**	4257

par:

$$\begin{aligned} \mathbf{K}^{hm}_{t|t+1} &= \mathbf{K}^{hm}_{t-1|t} + \eta \frac{\delta_{loss^{hm}}}{\delta_{\mathbf{K}^{hm}_{t-1|t}}} \\ \mathbf{B}^{hm}_{t|t+1} &= \mathbf{B}^{hm}_{t-1|t} + \eta \frac{\delta_{loss^{hm}}}{\delta_{\mathbf{B}^{hm}_{t-1|t}}} \end{aligned} \tag{7.2}$$

Évaluation de la performance de suivi

Afin de comparer de manière exhaustive les performances du système de suivi que nous proposons, nous avons évalué les performances de l'ensemble de données publiques—MOT17.

MOT17 contient 7 séquences vidéo de test pour évaluer les performances du tracker ; les informations des vidéos sont indiquées dans le tableau 5.1. Les scènes de MOT17 varient considérablement en termes d'arrière-plan, de conditions d'éclairage, de points de vue de la caméra et de mouvement de la caméra. Ces variations rendent MOT17 très difficile. Il convient de noter que les étiquettes et les boîtes de délimitation de la vérité de terrain des vidéos de test ne sont pas accessibles au public. L'évaluation est réalisée sur la base du benchmark MOTchallenge [2].

Notre tracker suit les vidéos de test sur la machine locale ; les résultats du suivi sont enregistrés dans des fichiers formatés en *.txt* et téléchargés sur le serveur hôte pour évaluation. Les résultats globaux de l'évaluation ont été présentés dans le tableau. 7.5. Nous avons résumé les résultats de l'évaluation des algorithmes de suivi actuels disponibles sur papier et en ligne sur le tableau de bord du serveur MOTchallenge. Comme nous l'avons vu dans les chapitres précédents, les performances du tracker, en particulier du MOTA, sont principalement déterminées par le réseau de détection. La méthode que nous proposons utilise

2. https://motchallenge.net

le même réseau de détection (même architecture de réseau et mêmes poids de réseau) que FariMOT; le MOTA est proche. Pour évaluer la fonction de chaque composant proposé, nous avons conçu deux implémentations. La première est le Scene-Adaptive Tracker, abrégé en SATracker. Elle se compose de deux éléments, dont un réseau de recherche de personnes - CenterNet-MDnet pour la détection des piétons et la représentation de la fonction de réidentification. Et le convertisseur adapté à la scène pour l'association de données. La seconde implémentation est Scene-Adaptive Tracker Plus, abrégée en SATracker+, une mise à jour de SATracker qui compose un module de détection supplémentaire adapté à la scène. La conception du module d'association adaptative à la scène consiste à convertir le problème d'association de données de l'association à l'échelle spatiale à l'association à l'espace des caractéristiques afin d'améliorer la capacité de suivi du tracker pour les scènes difficiles, c'est-à-dire les scènes en mouvement rapide. On peut le voir sur la table. 7.5 que le tracker SATracker a grandement amélioré la continuité du suivi tout en maintenant le MOTA à $73,12\%$. En comparant avec FairMOT, les cas d'IDSW ont diminué de moitié (de 3303 à 1569), et les fragments de piste ont également diminué pour un grand nombre d'entre eux (de 8073 à 4716). SATracker+ est équipé d'un module de détection adapté à la scène, conçu pour améliorer la détection des objets faibles, ce qui améliore encore la continuité du suivi. Les traces les plus perdues (ML) ont diminué de 375 à 294.

Conclusion

Ce chapitre propose un module de détection adapté à la scène. L'idée est d'utiliser les résultats du suivi pour renforcer la confiance du détecteur dans la détection d'objets positifs faibles. Nous proposons l'architecture globale du système de suivi et discutons en détail de la mise à jour de la liste de suivi. Sur la base du système de suivi proposé, nous obtenons des résultats de suivi sur l'ensemble de données de suivi difficiles—MOT17. Les résultats expérimentaux montrent que nos modules d'association et de détection adaptés à la scène sont efficaces pour améliorer les performances du tracker. L'application de l'association de caractéristiques de réidentification et l'amélioration de la détection améliorent considérablement la continuité du suivi du tracker. Tout au long de la conception, nous avons souligné l'importance de la rétroaction des informations. Les résultats de l'association sont renvoyés au convertisseur adapté à la scène pour en mettre à jour les poids. Les résultats du suivi sont renvoyés au réseau de détection pour mettre à jour les poids de la dernière couche convolutive. Nous pensons que l'"apprentissage" ne devrait pas exister uniquement dans la phase de formation. L'apprentissage en ligne de la scène actuelle est d'une importance capitale pour améliorer les performances de suivi.

Ronghua HU

Doctorat : Optimisation et Sûreté des Systèmes

Année 2021

Apprentissage profond de bout-en-bout pour la ré-identification et le suivi de personnes

Le suivi d'objets consiste à traiter de manière entièrement automatisée des séquences d'images et de vidéos pour la localisation objets d'intérêt et d'estimation de leurs trajectoires spatio-temporelles. En raison du changement d'apparence des personnes, de leurs mouvements non linéaires et de leurs occlusions mutuelles dans des scènes de foule, le suivi reste extrêmement complexe. Un système de suivi complet et robuste se compose d'un détecteur pour la détection sémantique, d'un réseau de ré-identification pour la représentation de l'apparence des piétons, et d'un module d'association pour la maintenance et la mise à jour des trajectoires. Dans cette thèse, on vise à intégrer les modules de détection en utilisant des technologies d'apprentissage profond pour la recherche et le suivi des piétons en mouvement par : (i) l'extraction automatique des caractéristiques d'apparence des personnes, (ii) la transformation adaptative des caractéristiques de ré-identification en caractéristiques d'association afin de prendre des décisions d'association sans la contrainte des boîtes englobantes, et (iii) la remontée en ligne des résultats de suivi vers le réseau de détection pour améliorer la détection des cibles moins résolues.

Mots clés : vision par ordinateur – détection du signal – apprentissage automatique – surveillance électronique.

Deep Learning End-to-end Person Search and Multiple Pedestrian Tracking

Multi-object tracking consists of fully automated processing of image and video sequences for locating objects of interest and estimating their Spatio-temporal motion trajectories. Thanks to the rapid development of deep learning technologies, multi-object tracking and detection are used in the field of safety and security. Due to the complex appearance changes of pedestrians, non-linear motion, and mutual occlusion in crowded and mobile scenes, multiple object tracking remain extremely complex and challenging. A complete and robust tracking system consists of a detector for semantic detection, a re-identification network for pedestrians' appearance embedding representation, and an association module for trajectory maintenance and updating. In this thesis, we aim to integrate these modules using deep learning technologies for multiple object tracking by: (i) proposing of a person search network, named FT-MDnet, to extract re-identification features from multiple types of mainstream detection networks that aims at the detection, localization, and matching of pedestrians on cross-camera image galleries, (ii) proposing of a scene adaptive data association module to convert re-identification features into association features for making association decisions without the constraint of bounding boxes, and (iii) proposing of a scene adaptive detection module online feeding back the tracking result to the detection network to enhance the detection of weak and small targets.

Keywords: computer vision – signal detection – machine learning – electronic monitoring.

Thèse réalisée en partenariat entre :

Ecole Doctorale "Sciences pour l'Ingénieur"