Person Re-Identification and its Application to Multi-Object Tracking

Vladimir Somers

Thesis submitted in partial fulfillment
of the requirements for the degree of
Ph.D. in Engineering Sciences

Dissertation committee:

Prof. Alexandre Alahi (EPFL, Switzerland, advisor)
Prof. Christophe De Vleeschouwer (UCLouvain, Belgium, advisor)
Prof. Laura Leal-Taixé (NVIDIA, Milano)
Prof. Andrea Cavallaro (EPFL, Switzerland)
Prof. Laurent Jacques (UCLouvain, Belgium)
Prof. Laurent Francis (UCLouvain, Belgium, chair)

Abstract

Multi-Object Tracking (MOT) is a fundamental computer vision task that involves detecting objects of interest in video frames and associating detections of the same object across time to form trajectories. For association, appearance cues, extracted through person re-identification (ReID) models, play a crucial role by capturing distinctive visual features of the tracked targets. However, despite its importance for tracking, ReID has primarily been studied as an image retrieval problem, with state-of-the-art methods overlooking tracking-specific challenges. This thesis focuses on advancing person re-identification methods, with a particular emphasis on making them more robust and better suited for tracking applications. A key challenge in ReID and MOT is handling occlusions, where targets become partially hidden by objects or other people, leading to degraded re-identification accuracy and potential identity switches in tracking. Additionally, tracking methods often fail to effectively combine ReID with motion cues and scene context, relying instead on naive association strategies. To address these challenges, this thesis makes three key contributions: BPBreID, a part-based method for robust occluded re-identification; KPR, a keypoint promptable ReID model designed to address multi-person occlusions scenarios; and CAMELTrack, an online tracking-by-detection method that replaces traditional heuristic for detection association with a context-aware learnable module. At the time of writing, KPR and CAMELTrack achieve state-of-the-art performance on widely-used benchmarks for occluded person re-identification and multi-object tracking. Complementary to this thesis work, additional research contributions were made in sports analytics, including the development of specialized re-identification models for athletes and the introduction of novel datasets for player tracking, re-identification, and jersey number recognition. The thesis concludes with a

critical analysis of the current state of tracking and re-identification technologies, offering my opinionated view about the future of these two rapidly evolving fields.

Keywords: *Deep Learning*, *Computer Vision*, *Multi-Object Tracking*, *Re-Identification*, *Supervised Learning*, *Deep Metric Learning*

Résumé

Le suivi d'objets multiples (MOT) est une tâche fondamentale de vision par ordinateur qui consiste à détecter des objets d'intérêt dans des images vidéo et à associer les détections d'un même objet à travers le temps pour former des trajectoires. Pour l'association, les indices d'apparence, extraits par les modèles de réidentification de personnes (ReID), jouent un rôle crucial en capturant les caractéristiques visuelles distinctives des cibles suivies. Cependant, malgré son importance pour le suivi, la ReID a principalement été étudiée comme un problème de recherche d'images, les méthodes à l'état de l'art négligeant les défis spécifiques liés au suivi. Cette thèse se concentre sur l'avancement des méthodes de réidentification de personnes, avec un accent particulier sur leur robustesse et leur adaptation aux applications de suivi. Un défi majeur en ReID et MOT est la gestion des occlusions, où les cibles deviennent partiellement cachées par des objets ou d'autres personnes, entraînant une dégradation de la précision de réidentification et des changements potentiels d'identité dans le suivi. De plus, les méthodes de suivi ne parviennent souvent pas à combiner efficacement la ReID avec les indices de mouvement et le contexte de la scène, s'appuyant plutôt sur des stratégies d'association naïves. Pour relever ces défis, cette thèse présente trois contributions clés : BPBreID, une méthode basée sur les parties de corps pour une réidentification robuste en cas d'occlusion ; KPR, un modèle de ReID guidé par points-clés conçu pour traiter les scénarios d'occlusions multi-personnes ; et CAMELTrack, une méthode de suivi en ligne par détection qui remplace les heuristiques traditionnelles d'association de détections par un module contextuel apprenable. Au moment de la rédaction, KPR et CAMELTrack font partie des méthodes les plus performantes sur les benchmarks largement utilisés pour la réidentification de personnes occultées et le suivi d'objets multiples. En complément de ces

travaux de thèse, des contributions de recherche additionnelles ont été réalisées dans le domaine de l'analyse sportive, comprenant le développement de modèles de ré-identification spécialisés pour les athlètes et l'introduction de nouveaux jeux de données pour le suivi des joueurs, la ré-identification, et la reconnaissance des numéros de maillot. Cette thèse se conclut par une analyse critique de l'état actuel des technologies de suivi et de réidentification, offrant mon opinion sur l'avenir de ces deux domaines en rapide évolution.

Acknowledgement

This thesis emerges from a unique collaboration between several institutions: a joint Ph.D. program between *UCLouvain* and *EPFL*, funded by *Keemotion*, a *UCLouvain* spin-off that was acquired by *Sportradar* during my research. While not officially part of the Ph.D. program, *ULiège* played a crucial role through our collaboration with *SoccerNet*, a non-profit organization that has created the most comprehensive open-source suite and benchmark for soccer video understanding. The journey that led to this thesis began five years ago, and like any meaningful adventure, it has been shaped by countless individuals who have supported, guided, and inspired me along the way. First and foremost, I want to thank my family. To my mother who revealed in me a passion for engineering and made pursuing engineering studies an obvious choice. To my father, who taught me the invaluable lesson of perseverance and following one's dreams. To my brother and sister, who have always been by my side and provided unwavering support these past years, even in the most challenging moments. To all my friends who encouraged me and had to endure the occasional "*I'm not available, I have a deadline coming up*" - I promise the deadlines are finally over! Thank you for the laughs, the encouragement, and for reminding me that life exists outside of research. I'm grateful to *ULiège* and the entire *SoccerNet* team, particularly *Anthony*, *Silvio*, and *Marc* - collaborating with you has been both an honor and a pleasure. Without your dedication to publishing new sports datasets, which are crucially needed in our field, I probably wouldn't have been able to bring the sports dimension to my Ph.D. At *EPFL*'s *VITA* lab, I had the privilege of working with an exceptional team - *Krishna*, *Kaouther*, *Bastien*, *Brian*, *Os*, *Reyhaneh*, *Po-Chien*, *Ahmad*, *Megh*, *Charles*, *Taylor*, *Kathrin*, *George*, *Saeed*, *Hossein* and many others who created an environment where ideas could flourish. Special thanks to *Carol* for making my one-year stay at *EPFL*

so seamless. At *UCLouvain*, thanks to all my colleagues: *Maxime I.*, *Gabriel*, *Niels*, *Abolfazl*, *Sarra*, *Alexandre*, *Antoine*, *Gilles*, *Dani*, *Maxime Z.*, *Benoit G.*, *Benoit B.*, *Anne-Sophie*, *Mathieu* and the others. Some of you I had the honor to collaborate with, and others I had the pleasure of taking very long *Geoguessr* breaks with. A special thanks to *Baptiste* and *Victor*, with whom we wrote two papers and open-sourced our *Tracklab* framework: thank you for the numerous discussions, debates, meetings, and late-night rushes before deadlines. The industrial perspective of this research wouldn't have been possible without *Sportradar*, formerly *Keemotion*. Thanks to all my colleagues *Marcello*, *Gaurav*, *Bastien*, *Jacek*, *Piotr*, *Milutin*, *Luka*, and many others - I'm looking forward to continuing our work together at *Sportradar*. A special thanks to *Cédric*, *Damien*, and *Pascaline* - this journey started with you a decade ago when I did my master's thesis at *Keemotion* under your supervision. Thank you *Alex Bustamante*, *Pitou*, and *Anne-Lise* for making this Ph.D. possible. *Davide*, your mentorship deserves particular recognition - your guidance, both technical and organizational, has been invaluable. To the members of my jury, thank you for your time and expertise in evaluating this thesis. Finally, to *Alex* and *Christophe*, thank you for your guidance and support throughout this process. Thank you for the numerous hours spent brainstorming and exchanging ideas - I have learned so much thanks to you, and I feel lucky and grateful to have had you as supervisors. Thank you for giving me the immense opportunity to pursue this Ph.D. I never imagined I would develop such a passion for research, and I don't regret a single moment spent working with you. This thesis represents not just my work, but the culmination of support, guidance, and collaboration from all of you. Thank you!

Contents

Abstract i

Résumé iii

Acknowledgement v

Contents vii

List of Figures xi

1 Introduction **1**
1.1 How to Read This Thesis 2
1.2 Core Research Questions and Contributions 3
1.2.1 Motivation and Research Questions 3
1.2.2 Core Contributions 5
1.3 Additional Contributions 6
1.3.1 Sports-related Technical Contributions 6
1.3.2 Datasets Contributions 8
1.3.3 Commitment to Open-source Research 9

2 Background **13**
2.1 Person Re-Identification 13
2.1.1 Problem Formulation 14
2.1.2 Strong Baseline for Holistic Re-Identification 16
2.1.3 Occluded ReID and Part-based Methods 18
2.1.4 Datasets 18
2.1.5 Evaluation Metrics 20
2.1.6 Re-Identification: Real-Life Applications and Societal Impact . 21

2.1.7 Useful Resources . . . 24
2.2 Multi-Object Tracking . . . 27
2.2.1 Problem Formulation . . . 27
2.2.2 Challenges in Multi-Object Tracking . . . 28
2.2.3 Multi-Object Tracking Paradigms . . . 28
2.2.4 Datasets . . . 29
2.2.5 Evaluation Metrics . . . 32
2.2.6 Useful Resources . . . 34

3 BPBreID . . . 37
3.1 Introduction . . . 38
3.2 Related Work . . . 40
3.3 Methodology . . . 41
3.3.1 Body Part Attention Module . . . 42
3.3.2 Global-local Representation Learning Module . . . 44
3.3.3 Overall Training Procedure . . . 45
3.3.4 Visibility-based Part-to-Part Matching . . . 47
3.4 Experiments . . . 48
3.4.1 Datasets and Evaluation Metrics . . . 48
3.4.2 Implementation Details . . . 48
3.4.3 Comparison with State-of-the-Art Methods . . . 50
3.4.4 Ablation Studies . . . 50
3.5 Conclusions . . . 55

4 KPR . . . 57
4.1 Introduction . . . 58
4.2 Related Work . . . 60
4.3 Methodology . . . 62
4.3.1 Tokenization . . . 62
4.3.2 MSF Swin Encoder . . . 64
4.3.3 Part-based Head . . . 64
4.3.4 Training Process . . . 65
4.3.5 BIPO Data Augmentation . . . 66
4.3.6 Inference . . . 66
4.3.7 KPR Pipeline Motivation and Intuition . . . 66
4.4 Experiments . . . 68
4.4.1 Evaluation Benchmarks and Metrics . . . 68
4.4.2 Proposed Occluded PoseTrack-ReID Dataset and Annotations 68
4.4.3 Implementation Details . . . 69

4.4.4 Comparison with State-of-the-Art Methods . . . 69
4.4.5 Multi-Person Pose Tracking in Videos . . . 71
4.4.6 Ablation Studies . . . 71
4.5 Conclusion . . . 74

5 CAMELTrack . . . 77
5.1 Introduction . . . 78
5.2 Related Work . . . 81
5.3 Methodology . . . 83
5.3.1 CAMELTrack Pipeline . . . 83
5.3.2 Our CAMEL Architecture . . . 85
5.3.3 Association-Centric Training . . . 87
5.4 Experiments . . . 88
5.4.1 Datasets and Metrics . . . 88
5.4.2 Implementation details . . . 88
5.4.3 Inference Speed . . . 89
5.4.4 Comparison to State-of-the-Art . . . 89
5.4.5 Ablation Studies . . . 91
5.4.6 Qualitative Analysis of Latent Representations . . . 94
5.4.7 Qualitative Results . . . 94
5.5 Conclusion . . . 97
5.6 Additional Discussion about MOT17 . . . 98
5.6.1 MOT17 Dataset Limitations . . . 98
5.6.2 Comparison with state-of-the-art methods . . . 99

6 Conclusion and Perspective . . . 101
6.1 Results Summary . . . 101
6.2 Limitations and Future Directions . . . 103
6.2.1 BPBreID . . . 103
6.2.2 KPR . . . 105
6.2.3 CAMELTrack . . . 105
6.3 My Opinionated View about the Future of ReID and MOT . . . 106
6.3.1 Person Re-Identification . . . 106
6.3.2 Multi-Object Tracking . . . 107
6.4 Final thoughts . . . 111

Bibliography . . . 115

List of Figures

1.1 Introduction: Contributions Diagram 12

2.1 Background: The Person Re-Identification Task 13
2.2 Background: Query-Gallery Person Retrieval 14
2.3 Background: Retrieval as a Ranking Task 14
2.4 Background: Representation Learning for ReID 15
2.5 Background: Market-1501 Leaderboard 16
2.6 Background: Bag-of-Tricks Architecture 17
2.7 Background: Training ReID Models 17
2.8 Background: Person Retrieval with Occlusions 18
2.9 Background: Global ReID Methods 19
2.10 Background: Part-based ReID Methods 21
2.11 Background: The Multi-Object Tracking Task 27
2.12 Background: Challenges for Long-term MOT 28
2.13 Background: Online vs Offline MOT 29
2.14 Background: Online Tracking-by-Detection Paradigm Diagram . 30
2.15 Background: MOT Taxonomy 31

3.1 BPBreID: Occluded-ReID Challenges 39
3.2 BPBreID: Model Architecture 43
3.3 BPBreID: Qualitative Results 54

4.1 KPR: Method Overview 59
4.2 KPR: The Multi-Person Ambiguity Issue 60
4.3 KPR: Model Architecture 63
4.4 KPR: BIPO Data Augmentation 67
4.5 KPR: Qualitative Pose Tracking Results 72

4.6 KPR: Qualitative Part-Attention Maps 74
4.7 KPR: Prompt Optionality Study 74
4.8 KPR: Qualitative Person Retrieval Results 75

5.1 CAMELTrack: Method Overview 78
5.2 CAMELTrack: Architecture Overview 84
5.3 CAMELTrack: Qualitative Tracking Results 95
5.4 CAMELTrack: Qualitative t-SNE Visualizations 96
5.5 CAMELTrack: Qualitative Similarity Distributions 96

1

Introduction

Multi-Object Tracking (MOT) is a fundamental computer vision task with wide-ranging applications across autonomous vehicles, robotics, human-computer interaction, biology, and *sports analytics*. At its core, MOT requires solving multiple computer vision tasks simultaneously: *detecting* objects in each frame, and modeling their *motion* and *appearance* to maintain their *identities* across time. Current state-of-the-art MOT systems address this challenge through pipelines that combine specialized models for each of these tasks. Within these tracking pipelines, *person re-identification* (ReID) plays a crucial role by capturing the distinctive appearance features of tracked targets, providing essential cues for solving the association problem across frames. However, despite its importance for tracking, person re-identification has primarily been studied as a *retrieval* problem, where the goal is to match query images against galleries of candidates across multiple non-overlapping camera viewpoints, mainly for street surveillance. This disconnect between ReID's development context and its practical deployment in tracking systems has led to several critical limitations in current approaches. This thesis focuses on advancing ReID capabilities specifically for *online human tracking* in monocular videos, where tracking decisions must be made immediately as each new frame arrives, without access to future information. We choose this online setting for several reasons: (i) it presents a more challenging scenario that better reflects human capabilities, (ii) it is crucial for live applications such as real-time sports analytics, and

(iii) it has emerged as the dominant paradigm in recent tracking research. Through a series of methodological innovations, we address three fundamental challenges in bridging the gap between ReID and tracking. First, we tackle the problem of *occlusions*, where traditional ReID models struggle when targets become partially hidden, leading tracking systems to discard potentially valuable appearance information. Second, we address the *multi-person ambiguity* problem that arises in crowded scenes, where standard ReID models lack mechanisms to identify their target person within bounding boxes containing multiple individuals. Finally, we examine the broader challenge of integrating ReID within tracking pipelines, developing novel approaches for *temporal feature aggregation*, *group-aware representation learning*, and *context-aware fusion* of motion and appearance cues.

This research was conducted as a joint PhD between *EPFL* and *UCLouvain*, in collaboration with *Sportradar*, a leading company in sports technology and innovation. While our core contributions advance the general state-of-the-art in ReID and tracking, we maintain a particular focus on solving challenges relevant to sports analytics, developing specialized player recognition systems and introducing new datasets to foster research in this domain. Finally, through my research journey, I have found that the majority of published work lacks accessible implementations and datasets. I believe this lack of accessibility holds back progress in this field, making it difficult to verify results or build upon prior work. As a result, I have made it a priority to keep this research *open* by sharing all code and datasets publicly.

1.1 How to Read This Thesis

This thesis primarily adapts and reformats work from three published papers, focusing on tracking and person re-identification (ReID). For readers new to these fields, I recommend starting with the "*Background*" section (Chapter 2) to understand the basic concepts and terminology before diving into the technical contributions. Readers already familiar with tracking and ReID may prefer to scan the core chapters (Chapter 3, Chapter 4, Chapter 5) while exploring "*Core Research Questions and Contributions*" (Section 1.2) and "*Additional Contributions*" (Section 1.3) to understand the overall motivation and contribution of this work. I also encourage reviewing "*Results Summary*" (Section 6.1) for an overview of our contributions, followed by "*Limitations and Future Directions*" (Section 6.2) and "*Future Perspectives on ReID and*

MOT" (Section 6.3), as these sections may fuel discussions about my work and the future of these research fields.

1.2 Core Research Questions and Contributions

This thesis makes three core contributions to the fields of person ReID and multi-object tracking. In this section, I present the key research questions that guided my work and the contributions that emerged from addressing them. All publications resulting from this research are summarized in Table 1.1 and Figure 1.1.

1.2.1 Motivation and Research Questions

I started this thesis with a clear objective in mind: advance research in re-identification by making these systems more robust to the challenges often encountered in multi-object tracking. During my first few months talking with my supervisors, colleagues and analysing the state-of-the-art, I discovered multiple key challenges that needed solving to properly integrate re-identification into tracking pipelines. I formulated these challenges into specific research questions that shaped my entire PhD journey and led to three main contributions. Let me walk you through these research questions - each one is tackled in detail in the main chapters of my thesis. A summary of the proposed solutions is provided hereafter and in the conclusion (Chapter 6).

1. **Occlusion-Robust Person Re-identification** [Chapter 3]: Occlusion presents a fundamental challenge for both re-identification and tracking systems. Current state-of-the-art tracking methods often discard re-identification cues when occlusions occur (typically indicated by low detection confidence scores). While this conservative approach aims to avoid unreliable matches, it discards potentially valuable appearance information that could aid tracking through occlusion. How can we develop re-identification methods that remain robust and reliable under partial occlusion scenarios?
2. **Multi-Person Ambiguity in Person Re-identification** [Chapter 4]: Current re-identification models typically operate on cropped bounding box images, assuming a single person of interest. However, this assumption breaks down in crowded scenes where bounding boxes

frequently contain multiple people due to occlusions. In these scenarios, the re-identification model has no way to determine which person within the bounding box is the intended target, leading to ambiguous and unreliable features precisely when they are most needed - during person interactions and crossings. While more sophisticated detection methods like pose estimation and instance segmentation explicitly resolve this ambiguity with pixel-level information, there is currently no established framework for re-identification models to leverage this fine-grained detection information. How can we resolve multi-person ambiguity in re-identification and establish effective communication between upstream detection systems and downstream re-identification models in tracking pipelines?

3. **Re-identification within Multi-Object Tracking Pipelines** [Chapter 5]: Online tracking presents unique challenges for re-identification that differ fundamentally from traditional person retrieval:

   (a) **Temporal aggregation of noisy ReID features**: Unlike image-based ReID that compares pairs of images, or video-based reid that compares short clean sequences, online tracking must compare a single detected instance (from the current frame) to a tracklet - a sequence of past observations that may contain noise like identity switches, occlusions, inaccurate bounding boxes, etc. How can we aggregate noisy detections temporally into a single robust tracklet-level representation?

   (b) **Group-aware Person Re-identification**: Unlike traditional ReID where (i) a person is compared to a large gallery of unknown candidates and (ii) features must capture absolute identity characteristics, tracking scenarios involve comparing against a reduced, known set of candidates - specifically the group of persons observed within the video. This provides an opportunity for relative feature learning, where representations can focus on discriminative characteristics within the observed group. How can we exploit this group context to increase the discriminativeness of each person's representation?

   (c) **Motion and ReID Cues Fusion**: When performing tracking in videos, two important cues are available for associating detections along time: motion and appearance. However, each cue's reliability varies significantly depending on the tracking context. For instance, appearance cues are crucial when targets leave

and reenter the camera field of view, while motion cues become essential when disambiguating between targets with similar appearances. Making an exhaustive list of all possible contextual scenarios affecting the discriminativeness of each cue is an impossible task. How can we develop a context-aware association strategy that dynamically balances motion and appearance cues in tracking? Moreover, how can this strategy generalize to incorporate arbitrary types of cues from different upstream models, such as jersey numbers in sports tracking, without requiring handcrafted rules or architectural modifications?

1.2.2 Core Contributions

We made three core technical contributions detailed in the three first chapters of this thesis, two about re-identification: BPBreID (WACV'23, Chapter 3) and KPR (ECCV'24, Chapter 4) and one about multi-object tracking: CAMELTrack (Under Review, Chapter 5). These contributions aimed at addressing the core research questions introduced above and led to three papers:

1. **BPBReID:** V. Somers, C. De Vleeschouwer, and A. Alahi, "Body part-based representation learning for occluded person re-identification," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, (Waikoloa, HI, USA), pp. 1613–1623, Inst. Electr. Electron. Eng. (IEEE), Jan. 2023

   Summary: We present BPBReID, a novel body part-based ReID model designed to handle partial occlusions. The model introduces two key innovations: (i) an attention module for predicting body part attention maps and extracting part-based features, and (ii) GiLt, a training loss that enables robust learning of part-based representations despite partial observability and non-discriminative local appearances. Our approach achieves state-of-the-art performance on occluded person re-identification.

2. **KPR:** V. Somers, A. Alahi, and C. D. Vleeschouwer, "Keypoint promptable re-identification," in *Eur. Conf. Comput. Vis. (ECCV)*, vol. 15137 of *Lect. Notes Comput. Sci.*, pp. 216–233, Springer Nat. Switz., Nov. 2024

   Summary: We introduce Keypoint Promptable ReID (KPR), a novel approach that addresses the Multi-Person Ambiguity problem in re-identification, where multiple individuals appear in a single bounding box. By reformulating re-identification as a promptable task using

semantic keypoints to specify the target person, our method enables unambiguous person matching in crowded scenes. To support this new paradigm, we contribute the Occluded PoseTrack-ReID dataset and provide keypoint annotations for four existing ReID benchmarks. Our approach demonstrates superior performance across multiple occluded person re-identification scenarios and extends effectively to pose tracking applications.

3. **CAMELTrack:** V. Somers*, B. Standaert*, V. Joos*, C. De Vleeschouwer, and A. Alahi, "Cameltrack: Context-aware multi-cue exploitation for online multi-object tracking," Under Review, Nov. 2023

   Summary: We present CAMELTrack, a novel tracking-by-detection framework for online multi-object tracking that maintains the modularity of traditional approaches while eliminating their reliance on hand-crafted heuristics. At its core is CAMEL, a Context-Aware Multi-Cue ExpLoitation module that learns robust association strategies directly from data using two transformer-based networks. Our approach remains modular, lightweight and efficiently trainable while leveraging external specialized models, achieving state-of-the-art performance across multiple tracking benchmarks.

1.3 Additional Contributions

While the core chapters focus on fundamental research in tracking and re-identification, my doctoral work also led to additional contributions in sports video understanding that are not detailed in this manuscript. Below, we outline these contributions and their related publications across three main categories. All these publications are also summarized in Table 1.1.

1.3.1 Sports-related Technical Contributions

Given the strong sports analytics focus of this PhD, we made two technical contributions specifically tailored to sports applications. We developed PRTreID (ACMW'23), a sports-oriented extension of BPBreID that addresses team sports-specific challenges in player re-identification, and introduced SoccerNet-GSR (CVPRW'24), a novel adaptation of multi-object tracking for sports that reconstructs the game state by tracking and identifying players on a minimap. Below, we detail these contributions and their related publications:

1. **SoccerNet-GSR:** V. Somers*, V. Joos*, S. Giancola, A. Cioppa, S. A. Ghasemzadeh, F. Magera, B. Standaert, A. M. Mansourian, X. Zhou, S. Kasaei, B. Ghanem, A. Alahi, M. Van Droogenbroeck, and C. De Vleeschouwer, "SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Seattle, WA, USA), Jun. 2024

   Summary: Traditional MOT systems track players using bounding boxes coordinates in the image space and assign them temporary identity labels that persist only during tracking. However, sports analytics demands more sophisticated capabilities: the system must localize players in court coordinate space and uniquely identify each tracked player with their real-world identity. Addressing these requirements, we introduce Game State Reconstruction (GSR), a novel computer vision task that aims to track and identify players from a single broadcast camera to reconstruct a video game-like minimap - all without requiring specialized hardware or player-worn sensors. Despite the significant value of such player tracking and identification system for the sports industry, no appropriate public benchmark existed until now. To address this need, we release SoccerNet-GSR, the first public benchmark for this task, comprising a dataset of 200 annotated video sequences, a novel evaluation metric called GS-HOTA, and an open-source baseline leveraging state-of-the-art deep learning methods. GS-HOTA is an extension of the HOTA metric, a popular evaluation metric for multi-object tracking, that accounts for the specificities of the GSR task, i.e. the additional target attributes (jersey number, role, team) and the detections provided as 2D points instead of bounding boxes.

2. **PRTreid:** V. Somers*, A. M. Mansourian*, C. De Vleeschouwer, and S. Kasaei, "Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking," in *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, ACM, Oct. 2023

   Summary: We present PRTreID, the first sports-specialized person re-identification model that jointly addresses three critical tasks related to team sports video analytics: player re-identification, role classification, and team affiliation. By employing multi-task learning on a shared backbone, we show quantitatively and qualitatively how PRTreid

creates richer, more discriminative part-based player representations. These representations can be used for standard re-identification, to group players by team using clustering approaches, and to predict athlete roles (player, goalkeeper, referee, etc). Our approach achieves state-of-the-art performance on both re-identification and multi-object tracking on the SoccerNet benchmark.

1.3.2 Datasets Contributions

When this thesis began, most publicly available tracking and re-identification datasets focused on street surveillance scenarios. Sports companies were reluctant to release their data due to high production costs, broadcasting rights restrictions, and fear of losing their competitive advantages. To address this critical gap in the sports video understanding community, we contributed four datasets, including the SoccerNet Game State Reconstruction dataset introduced above and three additional datasets. These datasets have served multiple sports computer vision challenges at CVSports (CVPR Workshop 2022-2024) and MMSports (ACM Workshop 2022-2023).

1. **SoccerNet-ReID:** S. Giancola, A. Cioppa, A. Deliège, F. Magera, V. Somers, L. Kang, X. Zhou, O. Barnich, C. De Vleeschouwer, A. Alahi, B. Ghanem, and e. a. Van Droogenbroeck, Marc, "SoccerNet 2022 challenges results," in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, ACM, Oct. 2022

   Summary: While person re-identification has been extensively studied in street surveillance scenarios, we introduce the first large-scale re-identification dataset specifically designed for sports. SoccerNet-v3 ReID focuses on matching soccer players across multiple camera views at the same time instant, presenting unique challenges absent from surveillance datasets: players from the same team share nearly identical appearances, identities have limited samples, and image resolutions vary significantly. The dataset contains 340,993 player thumbnails extracted from broadcast videos of 400 soccer games across 6 major leagues.

2. **SoccerNet-JN:** A. Cioppa, S. Giancola, V. Somers, F. Magera, X. Zhou, H. Mkhallati, A. Deliège, J. Held, C. Hinojosa, A. M. Mansourian, P. Miralles, O. Barnich, C. De Vleeschouwer, A. Alahi, B. Ghanem, and e. a. Van Droogenbroeck, Marc, "SoccerNet 2023 challenges results," *arXiv*, vol. abs/2309.06006, 2023

Summary: In this work, we introduce the first open dataset for tracklet-based jersey number recognition in soccer, a critical component for long-term player tracking.

3. **DeepSportRadar-ReID-v1:** G. Van Zandycke*, V. Somers*, M. Istasse*, C. D. Don, and D. Zambrano, "DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations," in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, ACM, Oct. 2022

   Summary: We release the first open-source re-identification dataset for basketball, focused on matching players from the same camera view across different timestamps. Derived from multi-object tracking video annotations, this dataset was specifically designed to improve tracking systems through better appearance-based player matching.

4. **DeepSportRadar-ReID-v2:** M. Istasse*, V. Somers*, P. Elancheliyan, J. De, and D. Zambrano, "DeepSportradar-v2: A multi-sport computer vision dataset for sport understandings," in *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, ACM, Oct. 2023

   Summary: Building upon DeepSportRadar-ReID-v1, we introduce an enhanced version featuring a broader range of challenging scenarios in basketball player re-identification. This version expands the dataset's scope and difficulty level, providing a more comprehensive benchmark for evaluating re-identification systems in basketball settings.

1.3.3 Commitment to Open-source Research

Throughout this PhD, I maintained a strong commitment to open research and reproducibility. To ensure transparency and accessibility, all codebases and datasets developed during my doctoral studies have been made publicly available via dedicated GitHub repositories, accessible on my profile: `https://github.com/VlSomers`.

Our re-identification research, including BPBreID, PRTreID, and KPR, was developed using the TorchReid framework[1]. Similarly, our tracking contributions, such as CAMELTrack and SoccerNet-GSR, were built upon

[1] `https://github.com/KaiyangZhou/deep-person-reid`

TrackLab[2], a multi-object tracking (MOT) research framework that I co-developed.

1. **Tracklab:** V. Joos*, V. Somers*, and B. Standaert*, "TrackLab." `https://github.com/TrackingLaboratory/tracklab`, 2024

   Summary: TrackLab is a modular and user-friendly framework tailored for multi-object tracking research. It supports various tracking paradigms, including pose estimation, segmentation, and bounding-box tracking, while accommodating diverse datasets and evaluation metrics. Notably, TrackLab serves as the primary codebase for the SoccerNet Game State Reconstruction challenge, hosted at `https://github.com/SoccerNet/sn-gamestate`.

2. **MaCVi MOT Competition:** B. Kiefer, M. Kristan, J. Pers, L. Zust, F. Poiesi, V. Somers, C. D. Vleeschouwer, A. Alahi, and F. A. de Alcantara Andrade et al., "1st workshop on maritime computer vision (macvi) 2023: Challenge results," in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, pp. 265–302, IEEE, 2023

   Summary: I achieved third place in the SeaDronesSee Multi-Object Tracking Competition at the 1st Workshop on Maritime Computer Vision (MaCVi) at WACV'23. My solution built upon BPBreID was presented orally at the workshop and detailed in this paper.

[2] `https://github.com/TrackingLaboratory/tracklab`

Id	Year	Nickname	Paper	Organism	(Co-) First Author	Topics					⭐	📚
						🧍	🎯	⚽	💾	🏆		
1	2023	BPBreID	**V. Somers**, C. De Vleeschouwer, and A. Alahi, *Body Part-Based Representation Learning for Occluded Person Re-Identification*, WACV'23.	EPFL + UCL	👍	✓					185	143
2	2024	KPR	**V. Somers**, A. Alahi, and C. De Vleeschouwer, *Keypoint Promptable Re-Identification*, ECCV'24.	EPFL + UCL	👍	✓	(✓)		✓		112	3
3	2024	CAMELTrack	**V. Somers***, B. Standaert*, V. Joos*, A. Alahi, and C. De Vleeschouwer, *Context-Aware Multi-cue ExpLoitation for Online Multi-Object Tracking*, (Under Review).	EPFL + UCL	👍	✓	✓				–	–
4	2024	SoccerNet GSR	**V. Somers***, V. Joos*, S. Giancola, A. Cioppa, ..., A. Alahi, and C. De Vleeschouwer, *SoccerNet Game State Reconstruction: End-to-End Athlete Tracking and Identification on a Minimap*, CVSports, CVPRW'24.	SoccerNet	👍		✓	✓	✓		274	19
5	2023	PRTreID	**V. Somers***, A. M. Mansourian*, C. De Vleeschouwer, and S. Kasaei, *Multi-task Learning for Joint Re-identification, Team Affiliation, and Role Classification for Sports Visual Tracking*, MMSports, ACM'23.	UCL	👍	✓	(✓)	✓			16	12
6	2023	MOT Competition	B. Kiefer, ..., **V. Somers**, C. De Vleeschouwer, and A. Alahi et al., *1st Workshop on Maritime Computer Vision (MaCVi) 2023: Challenge Results*, WACVW'23.	EPFL + UCL			✓			✓	–	27
7	2022	DeepSportradar-v1	G. Van Zandycke*, **V. Somers***, M. Istasse, C. Del Don, and D. Zambrano, *DeepSportradar-v1: Computer Vision Dataset for Sports Understanding with High Quality Annotations*, MMSports, ACM'23.	Sportradar	👍			✓	✓	✓	39	49
8	2023	DeepSportradar-v2	M. Istasse*, **V. Somers***, P. Elancheliyan, J. De, and D. Zambrano, *DeepSportradar-v2: A Multi-Sport Computer Vision Dataset for Sport Understandings*, MMSports, ACM'23.	Sportradar	👍			✓	✓	✓	–	13
9	2022	SoccerNet '22	S. Giancola, A. Cioppa, A. Deliège, F. Magera, and **V. Somers** et al., *SoccerNet 2022 Challenges Results*, MMSports, ACM'22.	SoccerNet				✓	✓	✓	52	39
10	2023	SoccerNet '23	A. Cioppa, S. Giancola, **V. Somers**, F. Magera, and X. Zhou et al., *SoccerNet 2023 Challenges Results*, arXiv'23.	SoccerNet				✓	✓	✓	19	24
11	2024	TrackLab	**V. Somers***, V. Joos*, and B. Standaert*, *TrackLab: Open-source Framework for Multi-Object Tracking*, GitHub repository, 2024.	EPFL + UCL	👍		✓				118	–
											815	**329**

Table 1.1 Summary of thesis contributions and their impact metrics. GitHub star count (⭐) and Google Scholar citations (📚) were captured on February 28th, 2025 at 10:30 CET. Nicknames and paper titles are clickable and link to the corresponding GitHub repository and publication respectively. Topics legend: 🧍 Person Re-identification, 🎯 Multi-Object Tracking, ⚽ Sports, 💾 New Dataset, 🏆 Workshop Competition.

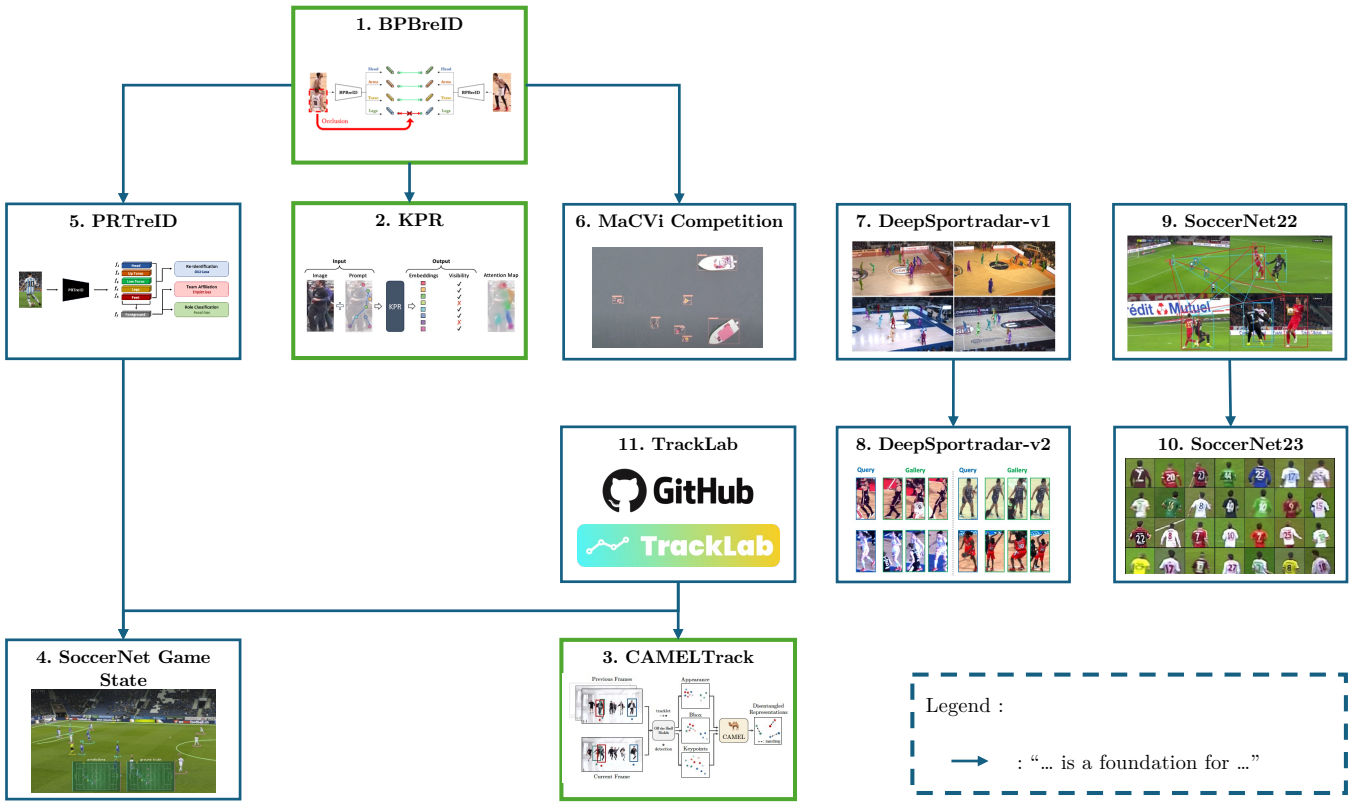


Fig. 1.1 A diagram illustrating the interdependency between my various contributions. The core contributions detailed in this thesis are highlighted with a green box.

2

Background

This chapter provides a concise introduction to person re-identification and multi-object tracking, focusing on their problem formulations, datasets, and evaluation metrics. We assume the reader has background knowledge in deep learning and computer vision. Prior literature is not extensively covered in this chapter, as each core section of the thesis contains its own related work discussion that contextualizes our contributions.

2.1 Person Re-Identification

This chapter provides a concise introduction to person re-identification (ReID), establishing key concepts essential for understanding the core contributions of this thesis. For comprehensive background, readers are directed to the excellent survey by [12]. Those interested specifically in occluded

Fig. 2.1 The core person re-identification objective.

Fig. 2.2 Person re-identification as a query-gallery retrieval task.

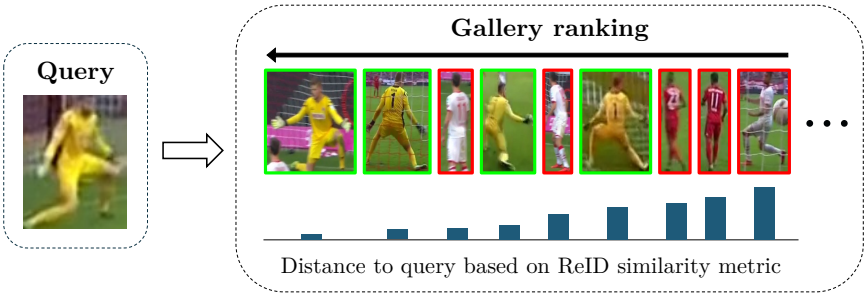


Fig. 2.3 Query-gallery retrieval by ranking the gallery.

re-identification, a central challenge addressed in this thesis, should consult the recent survey covering state-of-the-art approaches [13]. The chapter also examines the societal dimensions of ReID, clarifying common misconceptions and exploring its diverse applications beyond surveillance, from wildlife conservation to healthcare and sports analytics.

2.1.1 Problem Formulation

Person re-identification (ReID) addresses the fundamental challenge of determining whether two person images depict the same individual, regardless of when and where these images were captured. As illustrated in Figure 2.1, a ReID system can either make a binary decision about identity matching or, more commonly, compute a continuous similarity score that reflects the confidence of the match.

While ReID technology has various applications, from video tracking to 3D reconstruction, it is primarily studied in the context of person retrieval across camera networks. To better understand this task, we must first define two important concepts: the query and the gallery. The “query” is an image

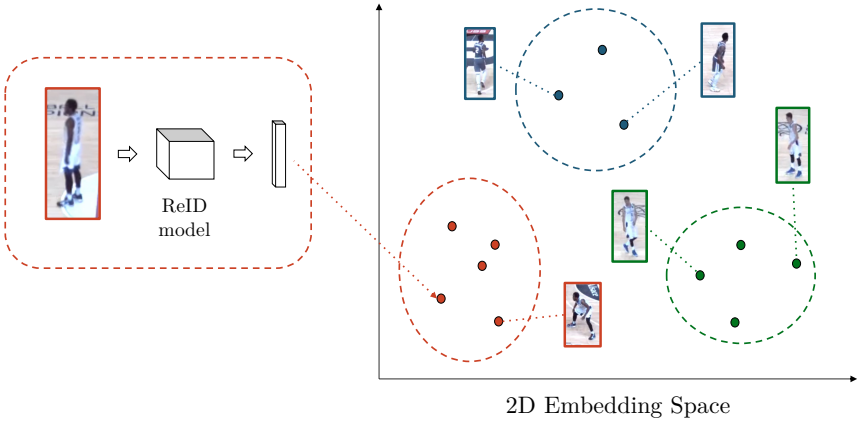


Fig. 2.4 Representation Learning for ReID.

of a person of interest captured from a specific camera that the ReID system is tasked with retrieving. The "gallery" is a large database of images of various individuals captured from different cameras at different points in time. For instance, all these images might be collected during one hour from a network of surveillance cameras in a city center [14].

In this setting, as shown in Figure 2.2, the task involves matching a query image against the gallery of candidate images. The goal is to retrieve all instances of the query person from within the gallery set.

This retrieval process, illustrated in Figure 2.3, is formulated as a ranking task where gallery images are ordered based on their similarity to the query. Given a query image, the ReID system computes its similarity to each gallery image and sorts them by decreasing similarity. The effectiveness of the system is measured by its ability to place true matches at the top of this ranked list.

Modern ReID approaches tackle this ranking problem through representation learning (Figure 2.4). These methods learn to map person images into a discriminative embedding space where images of the same person are positioned close together while images of different people are pushed apart. The similarity between any two images can then be quantified by measuring the distance between their respective embeddings, typically using cosine similarity.

Person ReID faces numerous challenges in real-world scenarios. Images of the same person often exhibit significant variations in viewpoint and

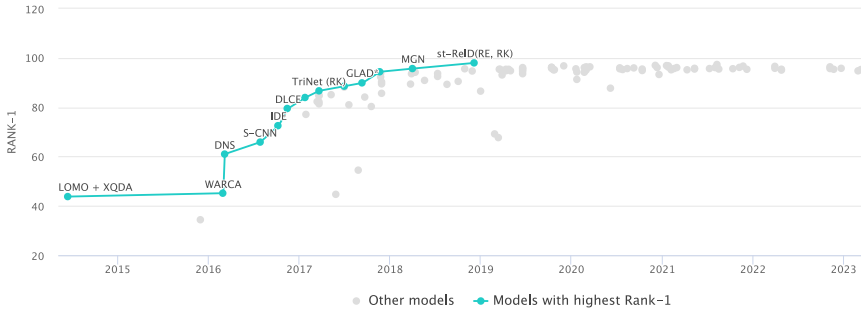


Fig. 2.5 Market-1501 Leaderboard as reported on PaperWithCode [1]. Performance metrics on this benchmark have plateaued since 2019.

pose across different cameras. The task is further complicated by varying illumination conditions, different camera characteristics, and the presence of background clutter. Poor image quality can also degrade matching performance.

Finally, occlusion represents one of the biggest obstacles for ReID models and will be discussed extensively throughout this thesis.

2.1.2 Strong Baseline for Holistic Re-Identification

The most commonly studied scenario in person ReID research is holistic re-identification, where datasets consist of full-body images with minimal occlusion. Market-1501 [14] remains the most widely used benchmark in this setting. As shown in Figure 2.5, performance metrics on this benchmark have largely plateaued since 2019, suggesting that holistic ReID has approached its performance ceiling.

The influential work "Bag of Tricks and a Strong Baseline" (nicknamed *BoT*) [15] serves as the foundation for many current ReID approaches, including ours. This work introduced several key techniques: using a ResNet-50 backbone with last stride removed, warm-up learning rate scheduling, random erasing augmentation, label smoothing regularization, and center loss. It employs the global ReID paradigm, as illustrated in Fig. 2.9.

The most significant contribution of *BoT* was its training approach combining two complementary objectives, as illustrated in Figure 2.7. The Identity loss uses cross-entropy to classify identities during training (discarded at inference), while the batch-hard triplet loss minimizes distances between same-identity pairs and maximizes distances between different

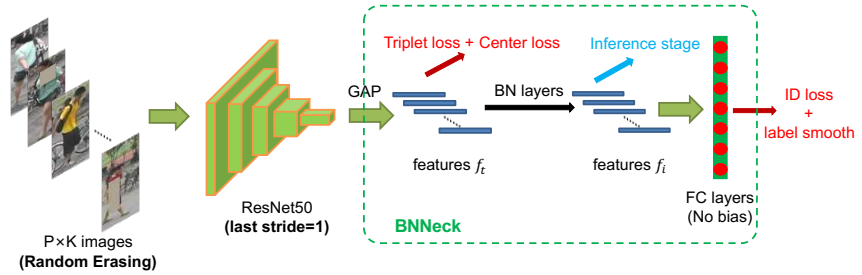


Fig. 2.6 Architecture of the *BoT* [15] model that serves as the baseline for numerous SotA ReID approaches, including ours.

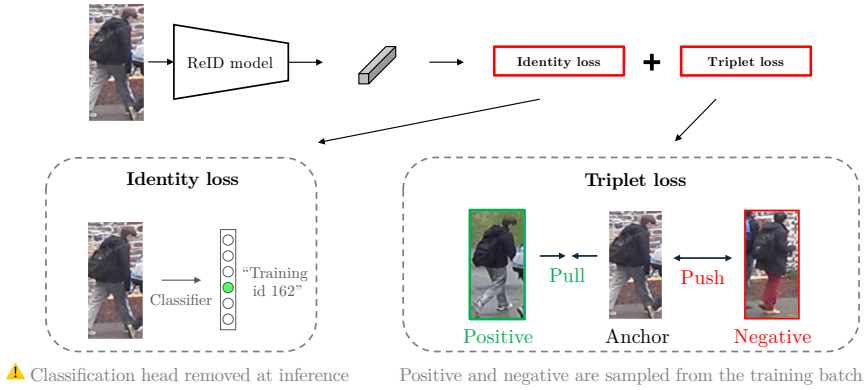


Fig. 2.7 Training ReID models with the Identity and Triplet losses.

identities. The combination of these two losses remains the dominant training paradigm in ReID.

While some papers [16] have claimed to reach human-level performance on holistic ReID benchmarks, research interest has shifted toward more challenging scenarios. These include cloth-changing ReID, domain adaptation, infrared ReID, sketch-based ReID, and particularly occluded ReID, which addresses real-world situations where subjects are partially visible. These emerging directions present significant challenges that remain largely unsolved, driving the next generation of ReID research.

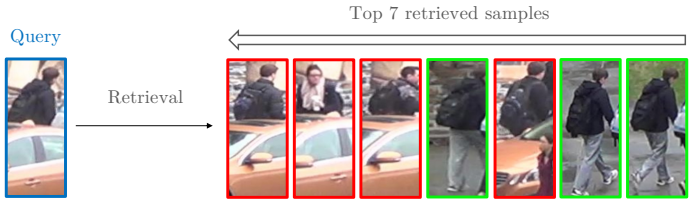


Fig. 2.8 Impact of occlusions on ReID retrieval performance.

2.1.3 Occluded ReID and Part-based Methods

Occlusion represents one of the most critical challenges for person re-identification and occurs frequently in multi-object tracking scenarios. As illustrated in Figure 2.8, occlusions significantly degrade retrieval performance by contaminating person representations with features from the occluding objects.

To address this challenge, specialized methods [17,18,18–26] employ the part-based paradigm. Part-based approaches decompose person images into distinct body regions and extract features from each part separately, as shown in Figure 2.10. These local features are then combined with global features to create more comprehensive representations with three key advantages: (1) they capture fine-grained appearance details that might be overlooked in global representations; (2) they enable explicit feature alignment that ensures corresponding body regions are directly compared across images; and (3) they allow for selective feature matching where only mutually visible parts are compared while occluded regions are disregarded.

We employ this part-based approach for our two main ReID contributions, which will be detailed further in Chapter 3 and Chapter 4. The fundamental benefits of part-based representation learning for handling partial observations are discussed more extensively in Section 6.2.1.

2.1.4 Datasets

Person re-identification research relies heavily on standardized datasets to benchmark methods. Most available ReID datasets are captured from camera networks across public places, in a typical street surveillance context. These datasets primarily focus on cross-camera matching, where individuals must be identified between images taken by different cameras, with query and gallery images coming from separate camera views.

It is important to note that test sets for most ReID datasets are pub-

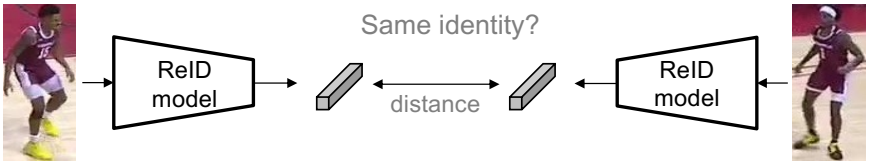


Fig. 2.9 Global ReID paradigm: Siamese networks outputs a single embedding per input image, that are then compared by their cosine distance.

licly available, which raises concerns about potential overfitting of the top-performing methods.

Standard ReID Datasets

Market-1501 [14] remains one of the most widely used benchmarks in the field. Another historically significant dataset, DukeMTMC-reID [27], has been withdrawn due to privacy concerns.

Occluded ReID Datasets

Our work primarily focuses on occluded person re-identification, for which several specialized datasets exist: Occluded-Duke [24], Partial-ReID [28], and Occluded-ReID [29]. A common characteristic of these datasets is that query samples are predominantly occluded, while the proportion of occluded samples in the training and galleries sets remains similar to standard ReID datasets.

Occluded-Duke stands as the most widely used dataset in this category and is notably the only one with a dedicated training set. Being derived from the withdrawn DukeMTMC-reID, we don't report performance on Occluded-Duke. It's important to note that most major conferences (CVPR, ECCV, ICCV, etc.) now prohibit papers from publishing results on this dataset. This situation has created a significant research dilemma, as the majority of occluded ReID methods were developed using a dataset that can no longer be featured in premier publication venues. The alternative occluded datasets (Partial-ReID and Occluded-ReID) are small and lack dedicated training sets, restricting their utility to model evaluation only. To address these limitations in the field, we introduce a new large-scale occluded dataset in our work KPR [2].

Dataset	Type	# Ids	# Cams	Train set	Occlusions
Market-1501 [14]	Street	1501	6	✓	
DukeMTMC-ReID [27]	Street	1812	8	✓	
MSMT17 [30]	Surveillance	4101	15	✓	
Occluded-Duke [24]	Street	1812	8	✓	✓
Occluded-ReID [29]	Outdoor	200	-		✓
Partial-REID [28]	Outdoor	60	-		✓
SoccerNet-ReID (ours) [6]	Soccer	243,432	-	✓	
Sportradar-ReID (ours) [8]	Basketball	486	-	✓	
Occ-PTrack (ours) [2]	Multi-sport	2411	-	✓	✓

Table 2.1 Comparing popular holistic and occluded re-identification datasets, including our proposed ones.

2.1.5 Evaluation Metrics

Person ReID systems are primarily evaluated using two complementary metrics: Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP). CMC-k measures the probability that a correct match appears within the top-k ranked gallery images. Most commonly, CMC-1 (also called Rank-1 accuracy) indicates how often the correct match is retrieved at the very top position. For a query i, the CMC-k accuracy is defined as:

$$\text{CMC-k}(i) = \begin{cases} 1, & \text{if a correct match appears in top-k positions} \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

The final CMC-k score is averaged over all N queries:

$$\text{CMC-k} = \frac{1}{N} \sum_{i=1}^{N} \text{CMC-k}(i) \tag{2.2}$$

While CMC is intuitive, it only considers the first correct match for each query and doesn't evaluate the model's ability to retrieve all instances of the person of interest. To address this limitation, mAP provides a more comprehensive evaluation by considering the retrieval of all correct matches. For each query i, the Average Precision (AP) is computed as:

$$\text{AP}(i) = \frac{1}{M_i} \sum_{j=1}^{M_i} \frac{j}{\text{rank}_j} \tag{2.3}$$

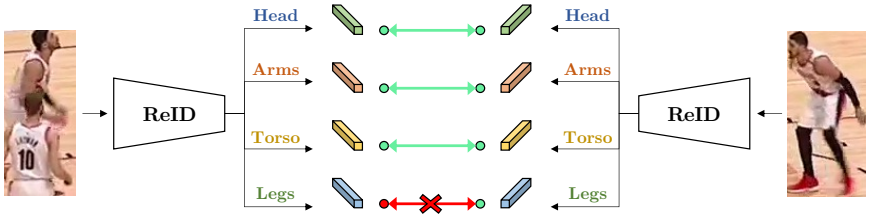


Fig. 2.10 Part-based ReID paradigm: Comparison is restricted to mutually visible body parts, with non-visible elements (legs in this example) excluded for computing a similarity score.

where M_i is the number of correct matches for query i in the gallery, and rank_j is the position of the j-th correct match in the ranked list. The mAP is then computed as the mean AP over all queries:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}(i) \tag{2.4}$$

Together, these metrics provide a balanced assessment of a ReID system's performance: CMC captures the crucial ability to find at least one correct match quickly, while mAP evaluates the completeness of the retrieval results.

2.1.6 Re-Identification: Real-Life Applications and Societal Impact

Person re-identification (re-ID) is often misunderstood, both in terms of its technical capabilities and its range of applications. This section first clarifies key misconceptions about the technology, particularly its distinction from identification systems, and then explores how re-ID has evolved beyond its surveillance origins to serve diverse sectors, from wildlife conservation to sports analytics.

Re-Identification is not Identification

Person "re-identification" should not be confused with person "identification". While the former aims at comparing two person images and answering the question "is it the same individual or not?", the latter aims at retrieving the exact identity of an individual based on any cue related to that person. Identification can be simply accomplished through the use of identification documents, such as a driver's license or passport, or with

more advanced techniques using biometric attributes, such as a fingerprint, facial features, or iris scan. Both tasks can be performed manually through human labor or automatically by an AI system. Although person re-ID can be used for many tasks, such as tracking an individual across a network of surveillance cameras, it cannot provide further personal information about the individual, such as his name or national identification number. Person re-identification serves therefore a different purpose than other person identification systems such as facial recognition.

Re-identification is not only street surveillance

Person re-identification systems were first developed for surveillance purposes and to improve public safety [14]. In this surveillance context, re-ID is used by authorities to track the movement and behavior of individuals in a given area, such as city centers, airports, train stations, or sports stadiums. The technology enables therefore human operators to easily detect and potentially prevent criminal or suspicious activity, identify potential suspects, or retrieve missing persons such as lost children. Finally, some countries are using these technologies to partially automate their social credit system [31], where businesses and individuals are evaluated for trustworthiness with a credit rating and are further blacklisted or whitelisted for specific services based on their personal score.

Although person re-identification was introduced in the context of mass surveillance in public places, it is actually a tool that has many other applications, most of which are unknown to the general public. Below, we provide a non-exhaustive list of such applications with short descriptions.

Wildlife protection Re-identification is not only limited to humans and has also been applied to animals. Such animal re-identification systems are very powerful tools used by NGOs for wildlife protection. For instance, re-identification systems are used to aid whale conservation efforts [32] and accelerate the counting and referencing of current populations. For the past 40 years, scientists have been using photo surveillance systems to monitor these whale populations, but manual photo identification and matching with existing databases was very timeconsuming. Consequently, a huge trove of data was left untapped and underutilized. With the use of whales' re-identification tools, scientists are now able to match newly captured images to known individuals faster. Such technology will revolutionize the scale at which long-term whale population data will be captured, which in turn will allow for a more detailed understanding of demographics,

social structure, reproductive rates, individual movement patterns, genetics, health, and causes of death.

Sports understanding and analytics Re-identification is used by sports federations and companies in a wide range of sports, including tennis, soccer, basketball, or ice hockey, to monitor players' performance, compute game analytics, assist video-assisting referees, and provide sports fans and amateurs with personalized content. For instance, re-identification has been used by companies [33] to provide amateur tennis players with automated video highlights of their games using a simple smartphone app. Re-identification is also used by sports tech companies [34] to provide basketball teams, leagues, and broadcasters with advanced analytics of each player's performance, such as distance traveled or shooting accuracy.

Robotics Following the recent advances of robotics, re-identification tools have enabled the development of a new type of assistive robots. These robots use the information from the ReID tool to determine the location and movements of the persons to assist and follow them accordingly. Commercial applications of these types of robots include automated golf caddies to carry equipment and follow the players [35], auto-follow camera drones for personalized self-video production [36] and wheeled robots to help elderly people carry their groceries [37].

Rescue Recently, re-identification systems have also been applied to maritime search and rescue (SaR) with drones [38]. Due to climate change and migration crises via major seas, maritime SaR has become increasingly important. Automated unmanned aerial vehicle are a very cheap, fast to deploy and risk-free solution to help finding distressed or missing vessel or person in the sea, and boat or swimmer re-identification software are mandatory for quickly localizing the rescue target.

Private sector Re-identification systems can also be deployed in private places [39], such as hospitals, nursing homes, and children's daycare, to ensure that only authorized individuals are able to enter certain areas of the facility (such as patient rooms, medication storage areas, or children's playground) or to keep track of patient location within the facility. For instance, these tools can help to prevent escape attempts from patients with moderate to severe dementia.

Key component for other vision tools Re-identification is also a key building block for other image and video analysis tools. For instance, tracking tools [40] are used to compute the trajectories of individuals within video sequences, and most existing tracking systems rely upon a re-identification software to gather important appearance cues about the tracked targets. Additionally, re-identification plays a crucial role in 3D reconstruction tools [41], which create three-dimensional images of people or objects from multiple two-dimensional camera perspectives. In these systems, re-identification is used to match the same individual across various camera views to reconstruct his pose. These 3D reconstruction tools cover a wide range of applications, including virtual and augmented reality or movie production.

Other industrial applications Finally, re-identification software can also be useful in a wide variety of domains in the industry, including transportation, agriculture, entertainment, retail or even military and defense, to track the movement and behavior of visitors, workers and passengers, to optimize the layout and operations in buildings, or to improve safety in general.

2.1.7 Useful Resources

The background section of this thesis provides only a starting point and overview of person re-identification. This section lists interesting resources for readers interested in delving deeper into the details or implementing their own ReID systems.

Awesome Lists

- **Awesome Re-ID with Leaderboard**: Collection of datasets with performance leaderboards, papers organized by year, and open-source frameworks.
  `https://github.com/FDU-VTS/Awesome-Person-Re-Identification`

- **Awesome Re-ID**: Curated list of person ReID papers organized by conference and subthemes, plus open-source implementations.
  `https://github.com/bismex/Awesome-person-re-identification`

- **Awesome Person Search**: Comprehensive resource for the related field of end-to-end person search.

https://github.com/daodaofr/Awesome-Person-Search

Benchmarks

- **Awesome Re-ID Datasets**: Tabulated summary of datasets with release dates, identities, image counts, and sample visualizations.
  https://github.com/NEU-Gou/awesome-reid-dataset

- **Performance Rankings**: Current state-of-the-art results on major benchmarks like Market-1501.
  https://paperswithcode.com/sota/person-re-identification-on-market-1501

Open-source ReID Frameworks

- **FastReID**: Efficient modular framework for state-of-the-art ReID models with comprehensive benchmarking.
  https://github.com/JDAI-CV/fast-reid

- **Torchreid**: PyTorch-based library for training and evaluating ReID models with an extensive model zoo.
  https://github.com/KaiyangZhou/deep-person-reid

- **OpenUnReID**: Open-source toolkit focused on unsupervised ReID methods.
  https://github.com/open-mmlab/OpenUnReID

Popular Papers

- **Person ReID Survey**: M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 2872–2893, Jun. 2022

- **Occluded ReID Survey**: Y. Peng, S. Hou, C. Cao, X. Liu, Y. Huang, and Z. He, "Deep learning-based occluded person re-identification: A survey," *arXiv*, vol. abs/2207.14452, 2022

- **Bag of Tricks and A Strong Baseline**: H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, (Long Beach, CA, USA), pp. 1487–1495, Inst. Electr. Electron. Eng. (IEEE), Jun. 2019

- **A very popular part-based ReID method**: T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *AAAI Conf. Artif. Intell.*, vol. 36, pp. 2540–2549, Association for the Advancement of Artificial Intelligence (AAAI), Jun. 2022
- **A pivotal paper about training ReID methods**: A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv*, vol. abs/1703.07737, 2017

2.2 Multi-Object Tracking

This section provides the necessary background on multi-object tracking (MOT) required to understand the core contributions of this thesis. We present key concepts, paradigms, and evaluation metrics that will be referenced throughout our work. For a comprehensive overview of the field, readers are directed to the survey by Wenhan Luo [43].

2.2.1 Problem Formulation

Multi-object tracking (MOT) aims to identify and track multiple objects throughout a video sequence. In its monocular form, which we focus on in this thesis, the input consists of a sequence of video frames, while the output is a set of tracks (or tracklets) $T_1, T_2, ..., T_m$. Each track T_m represents the temporal trajectory of a specific object, composed of a sequence of frame-level detections $\{d^1, d^2, ..., d^t\}$ where t indicates the frame number. This process is illustrated in Figure 2.11.

A detection d^t localizes a target object within frame t. This localization can take various forms depending on the application requirements: a bounding box, a set of keypoints representing a pose skeleton, or a pixel-

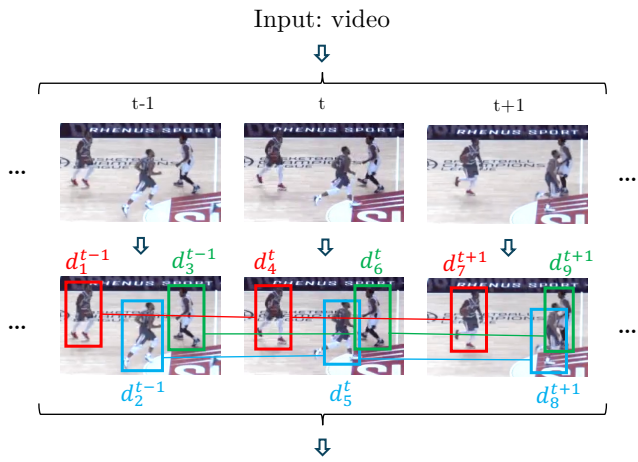


Fig. 2.11 Input and output of the Multi-Object Tracking (MOT) task. d_n^t refers to frame-level detections and T_m to the output tracks.

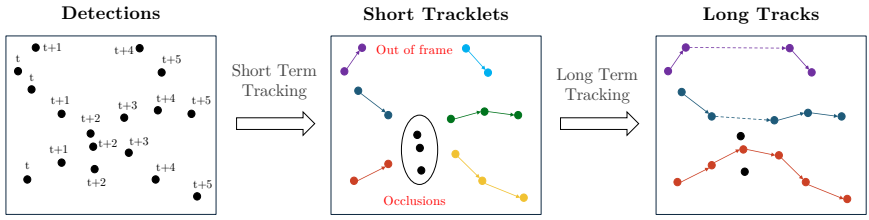


Fig. 2.12 Partial or total disappearance of tracked targets is a main challenge for long-term multi-object tracking: it occurs mainly during occlusions with scene objects or other tracked targets, or when targets exit and re-enter the camera frame.

wise segmentation mask. Most challenging public MOT benchmarks focus on bounding box tracking, making it the primary research focus of the community. In this thesis, we explore tracking methods that operate on both bounding boxes and keypoints.

2.2.2 Challenges in Multi-Object Tracking

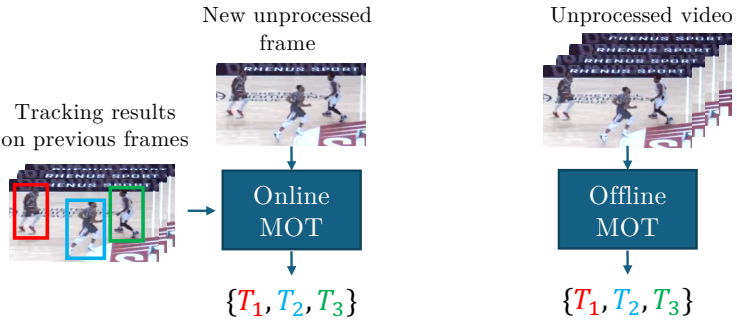
With recent advances in object detectors, short-term tracking has become increasingly reliable. However, significant challenges persist when targets partially or completely disappear from view, as illustrated in Figure 2.12. These disappearances occur primarily during occlusions (either between objects or with scene elements) and when targets exit and re-enter the camera frame.

2.2.3 Multi-Object Tracking Paradigms

Figure 2.15 provides a comprehensive overview of the current important paradigms in MOT together with their definitions, and categorizes popular recent methods in the field. This taxonomy organizes approaches along two primary dimensions: temporal processing requirements (online vs. offline) and tracking strategies (Tracking-by-Detection vs. Detection-by-Tracking). A visual representation of the offline and online paradigms is further illustrated in Figure 2.13. This thesis focuses primarily on online tracking due to its applicability to real-time sports analytics.

Online Tracking-by-Detection

We briefly describe the Online Tracking-by-Detection (TbD) paradigm as it represents the most widely adopted approach in current MOT research,



(a) Online **(b)** Offline

Fig. 2.13 Online vs Offline MOT paradigms.

consistently achieving state-of-the-art performance across major benchmarks while maintaining an intuitive and modular design. As illustrated in Figure 2.14, this approach follows a two-stage process: detection and association. First, an off-the-shelf object detector is applied on each video frame individually. Second, the association process involves: (1) cue extraction (motion, appearance, etc.) to build a tracklet-detection cost matrix, (2) optimization (e.g., Hungarian algorithm) to compute the optimal matching, and (3) track management to update matched tracks, initiate new ones, or terminate obsolete tracks.

2.2.4 Datasets

Creating multi-object tracking datasets is particularly resource-intensive, requiring frame-by-frame annotation of object identities across video sequences. This laborious process explains the historical scarcity of large-scale MOT datasets.

MOTChallenge [44] has been the dominant benchmark in the field for many years, consisting of the MOT15, MOT16, MOT17, and MOT20 datasets. Despite its widespread adoption, MOTChallenge has notable limitations: the dataset is relatively small, lacks a proper validation set for conducting experiments, and primarily features pedestrians in surveillance scenarios with predominantly linear motion patterns.

Recently, several new datasets have emerged to address these limitations and present novel challenges:

DanceTrack [45] focuses on tracking dancers with similar appearances

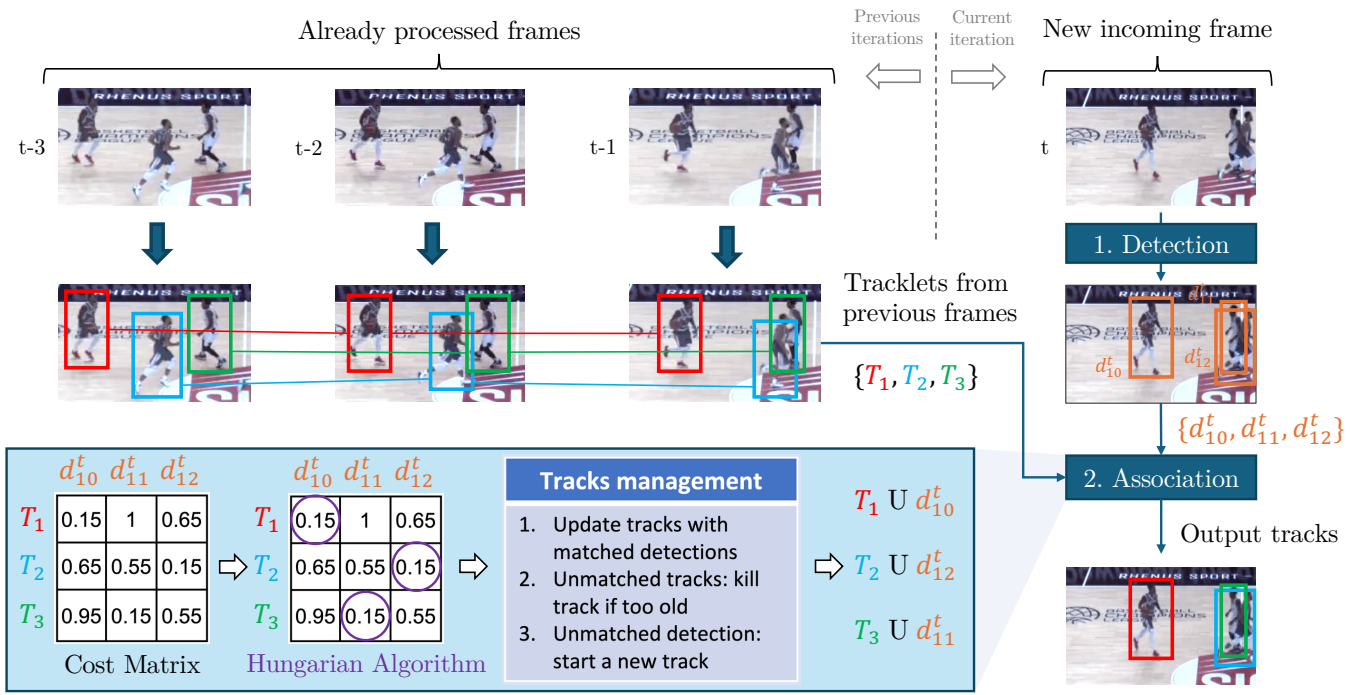


Fig. 2.14 Online Tracking-by-Detection Paradigm: A two-stage approach comprising a first detection step in the current image frame, followed by an association step with previously built tracklets.

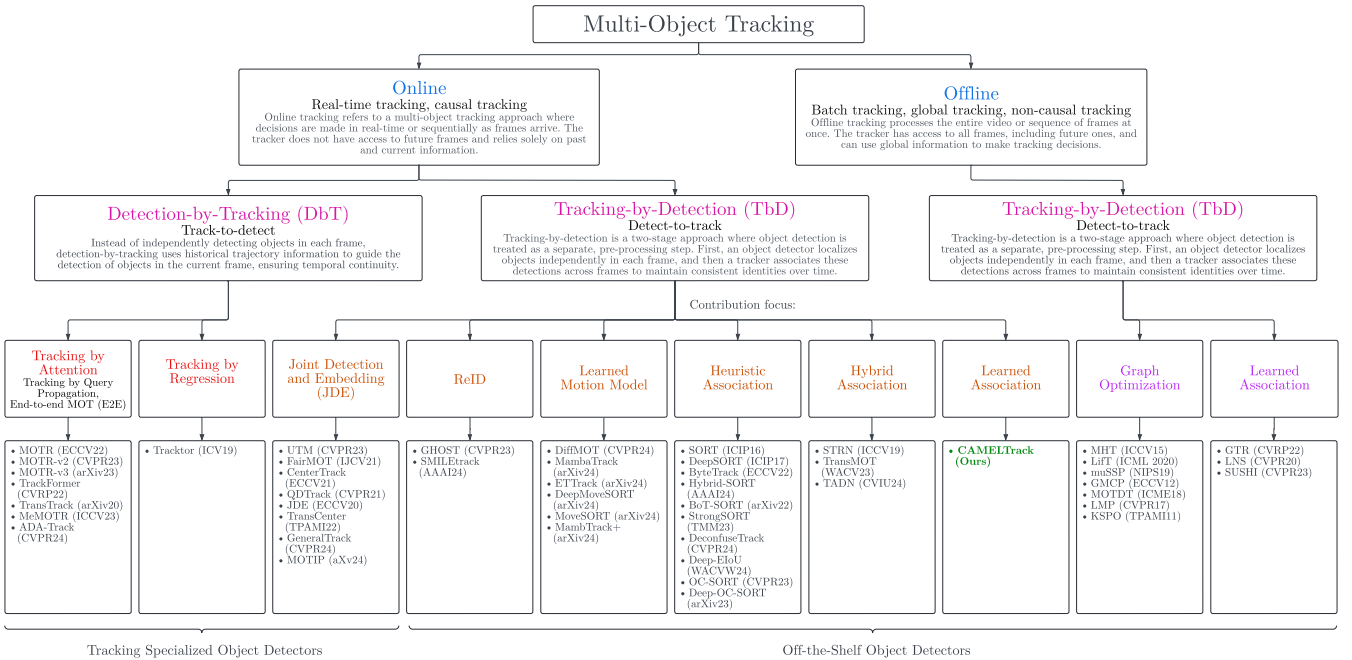


Fig. 2.15 Taxonomy of Multi-Object Tracking (MOT) approaches. Categorizing methods is challenging as some span multiple groups. This taxonomy emphasizes one of our core contributions: CAMELTrack.

executing complex, non-linear movements with many occlusions. This dataset specifically challenges appearance-based association methods and encourages more sophisticated motion modeling.

SportsMOT [46] features fast-moving athletes in soccer, basketball, and volleyball. It presents frequent occlusions, rapid direction changes, and scene re-entries that challenge re-identification capabilities.

SoccerNet Tracking [47] specializes in soccer player tracking with challenges similar to SportsMOT, but focuses exclusively on soccer and provides additional annotations such as jersey number, team, and player roles.

PoseTrack21 [48] expands tracking to human pose estimation across frames, requiring systems to track individual keypoints rather than just bounding boxes, adding an extra dimension of complexity. It primarily features handheld amateur videos with various sports activities.

These newer datasets, with their larger training sets and diverse scenarios, are encouraging research into previously underexplored challenges in MOT, including non-linear motion, appearance similarity, rapid movement patterns, strong occlusions, and scene re-entries. A complete comparison is provided in Table 2.2.

Table 2.2 Comparison of Multi-Object Tracking Datasets

Dataset	Videos	Frames	Tracks	Boxes
MOT17	14	11,235	1,342	292,733
MOT20	8	13,410	3,456	1,652,040
DanceTrack	100	105,855	990	–
SportsMOT	240	150,379	–	1,629,490
SoccerNet	201	225,375	5,009	3,645,661
PoseTrack21	593	89,000	–	–

2.2.5 Evaluation Metrics

Evaluating multi-object tracking performance requires metrics that assess both localization accuracy and identity consistency across frames. We present an overview of the three dominant metric families used in MOT benchmarks: CLEAR MOT metrics [49], IDF1 [50], and HOTA [51].

Before computing the evaluation metric, the detected objects must be matched with ground truth objects using the Hungarian algorithm, which minimizes the global matching cost based on Intersection-over-Union (IoU):

$$\text{IoU}(A, B) = \frac{A \cap B}{A \cup B} \tag{2.5}$$

Only matches with IoU above a certain threshold are considered valid, using either a fixed threshold or integrating across multiple thresholds.

CLEAR MOT Metrics

The CLEAR MOT metrics [49] established standardized tracking evaluation. Multiple Object Tracking Accuracy (MOTA) quantifies tracking errors relative to ground truth:

$$\text{MOTA} = 1 - \frac{|FN| + |FP| + |IDSW|}{|\text{gtDet}|} \tag{2.6}$$

where $IDSW$ are identity switches and gtDet is the total number of ground truth objects.

MOTA is heavily biased toward detection performance (FP and FN), with association errors (IDSW) typically much fewer in number despite their importance for tracking quality.

IDF1 Metric

IDF1 [50] matches at the trajectory level rather than detection level. It calculates ID-Precision and ID-Recall using Identity True Positives (IDTP), Identity False Negatives (IDFN), and Identity False Positives (IDFP):

$$\text{IDF1} = \frac{2|\text{IDTP}|}{2|\text{IDTP}| + |\text{IDFN}| + |\text{IDFP}|} \tag{2.7}$$

While IDF1 better captures association performance, it has a strong bias toward association quality.

HOTA Metric

The Higher Order Tracking Accuracy (HOTA) metric [51] provides a balanced assessment of detection and association:

$$\text{HOTA} = \sqrt{\text{DetA} \cdot \text{AssA}} \tag{2.8}$$

HOTA balances detection (DetA) and association (AssA) performance, ensuring improvements in either aspect are properly reflected in the final score.

Many benchmarks now report all three families of metrics, with HOTA increasingly adopted as the primary evaluation measure for new MOT

datasets.

2.2.6 Useful Resources

This section compiles useful resources for readers interested in delving deeper into the details or implementing their own MOT systems.

Awesome Lists

- **Multi-Object Tracking Paper List**: Systematically organized collection of papers by methodology (traditional methods, deep learning methods) and application scenarios.
  `https://github.com/SpyderXu/multi-object-tracking-paper-list`

- **Deep Learning for Tracking and Detection**: Comprehensive resource covering both tracking and re-identification aspects.
  `https://github.com/abhineet123/Deep-Learning-for-Tracking-and-Detection?tab=readme-ov-file#re_id_`

- **Awesome Multi-Object Tracking**: Comprehensive collection of papers, datasets, metrics, and code implementations for MOT.
  `https://github.com/luanshiyinyang/awesome-multiple-object-tracking`

Open-source MOT Frameworks

- **BoxMOT**: Collection of state-of-the-art tracking methods with unified implementation.
  `https://github.com/mikel-brostrom/boxmot`

- **TrackEval**: Framework for standardized evaluation of multiple object tracking algorithms.
  `https://github.com/JonathonLuiten/TrackEval`

- **ByteTrack**: Simple and effective framework achieving SOTA performance without complicated association techniques.
  `https://github.com/ifzhang/ByteTrack`

- **DeepSORT**: The most popular tracking method.
  `https://github.com/nwojke/deep_sort`

Leaderboards

- **MOTChallenge**: Standard benchmark evaluation platform for multi-object tracking.
  `https://motchallenge.net/`

- **Specialized MOT Datasets**: Leaderboards for DanceTrack, SportsMOT, and SoccerNet tracking challenges.
  `https://codalab.lisn.upsaclay.fr/`

Popular Papers

- **Multiple object tracking: A literature review**: W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T. Kim, "Multiple object tracking: A literature review," *Artif. Intell.*, vol. 293, p. 103448, 2021

- **DeepSORT: Simple Online and Realtime Tracking with a Deep Association Metric**: N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE Int. Conf. Image Process. (ICIP)*, (Beijing, China), pp. 3645–3649, Inst. Electr. Electron. Eng. (IEEE), Sept. 2017

- **Tracktor: Tracking without bells and whistles**: P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 941–951, Inst. Electr. Electron. Eng. (IEEE), Oct. 2019

- **MOTR: End-to-End Multiple-Object Tracking with Transformer**: F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "Motr: End-to-end multiple-object tracking withătransformer," in *Eur. Conf. Comput. Vis. (ECCV)*, (Cham), pp. 659–675, Springer Nat. Switz., 2022

3

BPBreID: Body Part-Based Representation Learning for Occluded Person Re-Identification

Published at WACV 2023

3.1 Introduction

Person re-identification [12, 13], or ReID, is a person retrieval task which aims at matching an image of a person-of-interest, called the query, with other person images from a large database, called the gallery. ReID has important applications in smart cities for video-surveillance [14, 27] or sport understanding [6, 8]. Person re-identification is generally formulated as a representation learning task and is very challenging, because person images generally suffer from background clutter, inaccurate bounding boxes, pose variations, luminosity changes, poor image quality, and occlusions [13] from street objects or other people.

For solving the ReID task, most methods adopt a global approach [15, 42], learning a global representation of the target person as a single feature vector. However, these methods are unable to address the challenges caused by occlusions for two reasons, both depicted in Figure 3.1:

⟨1⟩ *Obstacle and pedestrians interference*: the globally learned representation might include misleading appearance information from occluding objects and pedestrians.

⟨2⟩ *Requirement for part-to-part matching*: When comparing two occluded samples, it is only relevant to compare body parts that are visible in both images. Global method cannot achieve such part-to-part matching, because the same global feature is used for every comparison.

To deal with the above issues, part-based approaches [20, 26, 54], have shown promising results. These part-based methods address the ReID task by producing multiple local feature vectors, i.e., one for each part of the input sample. However, learning such part-based representations involves dealing with two crucial challenges:

⟨3⟩ *Non-discriminative local appearance*: Standard ReID losses, such as the id or triplet losses, work with the assumption that different identities have different appearance, and consequently that their corresponding global feature vectors are different. However, this assumption is broken when working with part-based feature vectors, because two persons with different identities might have very similar appearance on some of their body parts, as depicted in Figure 3.1. Because local appearance is not necessarily discriminative, standard ReID losses used for learning global representations do not scale well to local representation learning. The specificity associated to learning local features and its impact on the choice of the training loss has been overlooked in previous part-based ReID works and we are the first to point it out. To address these issues, we propose *GiLt*, a novel training

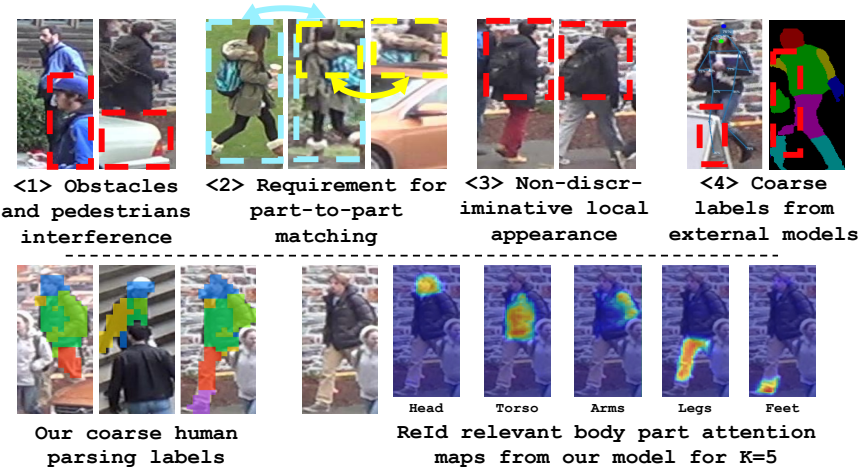


Fig. 3.1 Overview of key concepts in our work. First row illustrates the four challenges of occluded and part-based ReID that our proposed method is trying to address. Second row illustrates our pre-generated human parsing labels and the ReID-relevant soft attention maps produced by our model BPBreID.

loss for part-based methods. GiLt is designed to be robust to occlusions and non-discriminative local appearance, and is meant learn a set of local features that are each representative of their corresponding local parts, while being discriminative when considered jointly.

⟨4⟩ *Absence of human topology annotation*: Part-based methods generally rely on spatial attention maps to perform local pooling within a global feature map and build body part features of the ReID target. However, no ReID dataset is provided with annotations regarding the local region to pool, and generating such annotation with external pose information or part segmentation tools yields inaccurate results due to the domain variation and poor image quality. Moreover, body part-based feature pooling fundamentally differ from pixel-accurate human parsing. Indeed, the spatial attention maps have to localize the body part in the image, but also to identify the feature vectors that best represent discriminant characteristics of the body part appearance. Therefore, an ideal attention map is not necessarily an accurate segmentation shape. Previous ReID works exploiting human parsing to build part-based features have either (i) used directly the output of a pose estimation model as local attention masks, without adapting it to

handle the ReID task [17, 19, 24], or (ii) learned local features with part discovery, without human topology prior [20, 23, 55]. In this work, we propose a body part attention module trained with a novel dual supervision, using both identity and coarse human parsing labels. This module demonstrates how external human semantic information can be effectively leveraged to produce ReID-relevant body part attention maps.

Finally, we combine this body part attention module and the GiLt (Global-identity Local-triplet) loss to build our Body Part-Based ReID model called **BPBreID**, which effectively addresses all four challenges introduced before. We summarize the main contributions of our work as follows:

1. For the ReID task, we are the first to propose a soft attention trained from a dual supervision, to leverage both identity and prior human topology information. Our work demonstrates that this approach outperforms all previous part-based methods.
2. We propose a novel *GiLt* strategy for training part-based method. GiLt is robust to occlusions and non-discriminative local appearance and is straightforward to integrate with any part-based framework.
3. BPBreID achieves state-of-the-art performance on the occluded ReID task. Our BPBreID codebase has been released to encourage further research on part-based methods.

3.2 Related Work

Part-based feature alignment in ReID: To solve the spatial misalignment issue, several works [17, 19, 24, 24, 26, 56–61] adopt fixed attention mechanisms, using pre-determined pixel partitions of the input image and applying part pooling for generating local feature representations. These methods achieve poor feature selection and alignment, because the resulting attention maps are not meant for pooling ReID-relevant body part features. To solve those issues, other works [18, 20, 62, 63] use attention mechanisms trained in an end-to-end fashion for generating attention maps that are specialized towards solving the ReID task. Some of these approaches [18, 62, 63] include a pose estimation backbone as a parallel branch, which is jointly trained with the appearance backbone branch in an end-to-end fashion on the ReID dataset. However, the parallel branch induces a significant computational overhead and these methods do not address the occluded ReID problem explicitly. Other part-based methods learn local features via part discovery in a self-supervised way, without human topology prior [20, 23, 55]. Such

approach might introduce alignment errors, missed parts and background clutter. Different from previous works, our part-based features are built by an attention branch which (i) is trained explicitly to pool local features that are relevant for the ReID task, and (ii) leverages external human parsing labels to bias the spatial attention in focusing on prior body regions.

Local feature learning in ReID: Identity loss and batch-hard triplet loss [42] are two popular objectives for training ReID models that are also applied to part-based methods [17, 19, 20, 24, 25, 61, 64, 65] for learning local representations. Most of these methods [24,26,64,66] solely apply an identity loss on each part-based features. As a consequence, they are more sensitive to non-discriminative body parts and miss out the benefits [15, 42] of the triplet loss as a complementary deep metric learning objective for ReID. To deal with incomplete information of part features, some works [20, 65] propose to apply triplet and identity losses on a combined embedding, resulting from concatenation or summation of local features. A specialized Improved Hard Triplet Loss (IHTL) is proposed in [65] for training part-based feature, but this objective cannot cope well with occluded or similar samples. Finally, [25] applies both triplet and identity losses on combined part features, but does not use holistic features during training, which renders their training scheme less robust to inaccurate body part predictions and heavy occlusions. Our proposed GiLt training procedure aim at solving above issues and addressing the lack of consensus regarding the choice of losses to adopt for training part-based methods. Finally, it is worth noting that other works [19, 23] take an opposite approach to deal with standard ReID losses being unsuitable for non-discriminative body parts appearance. They solve this by constraining each part-based feature to be discriminative on its own, by either having each of them attending simultaneously to multiple body regions [23] or by adding high-order information in each local feature via message-passing [19].

3.3 Methodology

The overall architecture of our model BPBreID is depicted in Figure 3.2. It comprises two modules: the body part attention module described in Section 3.3.1 and the global-local representation learning module described in Section 3.3.2. The overall training procedure of BPBreID is described in Section 3.3.3 and the procedure used at inference for computing query to gallery distance is described in Section 3.3.4.

3.3.1 Body Part Attention Module

The body part attention module takes as an input the feature map extracted by the backbone and outputs a set of attention maps highlighting the body parts of the ReID target. This module consists of a *pixel-wise part classifier* trained with a *body part attention loss* using our coarse *human parsing labels*. We detail these three components below. Because our model is trained end-to-end, the body part attention module also receives a training signal from the ReID loss, which uses the identity labels, as described in Section 3.3.3. This attention branch is therefore trained from a **dual supervision**, with both a body part prediction objective and a ReID objective. As a result of the dual supervision, this module generates attention maps that are more relevant to the ReID task than the attention maps we would obtain using the fixed output of a pre-trained human parsing model. This module is depicted in the top left part of Figure 3.2.

Pixel-wise Part Classifier

The body part attention module takes in input the appearance map G, which is a tensor $R^{H \times W \times C}$ produced by a feature extractor. For each pixel (w, h) in the appearance map G, a pixel-wise part classifier predicts if it belongs to the background or to one of the K body parts, which means there are $K+1$ target classes, with the class at index 0 being the background. A 1x1 convolution layer with parameters $P \in R^{(K+1) \times C}$ followed by a $softmax$ is applied on G to obtain the classification scores $M \in R^{H \times W \times (K+1)}$:

$$M = \text{softmax}(GP^T)\ . \tag{3.1}$$

These $K+1$ probability maps M_k indicates therefore which pixels belong to which body parts (or to the background).

Human Parsing Labels

Human parsing labels $Y \in R^{H \times W}$, required for training our part attention module, are generated with the PifPaf [67] pose estimation model. $Y(h, w)$ is set to {1, ..., K} if spatial location (h, w) belong to one of the K body parts or 0 for background. Human semantic regions are defined manually for a given value of K. For instance, with $K = 8$, we define the following semantic regions: {*head*, *left/right arm*, *torso*, *left/right leg* and *left/right feet}*. These coarse human semantic parsing labels are illustrated in Figure 3.1 for $K = 5$. BPBreID could be easily applied to other domains, like car or animal re-identification, by replacing these human parsing labels with other parsing labels relevant to the domain.

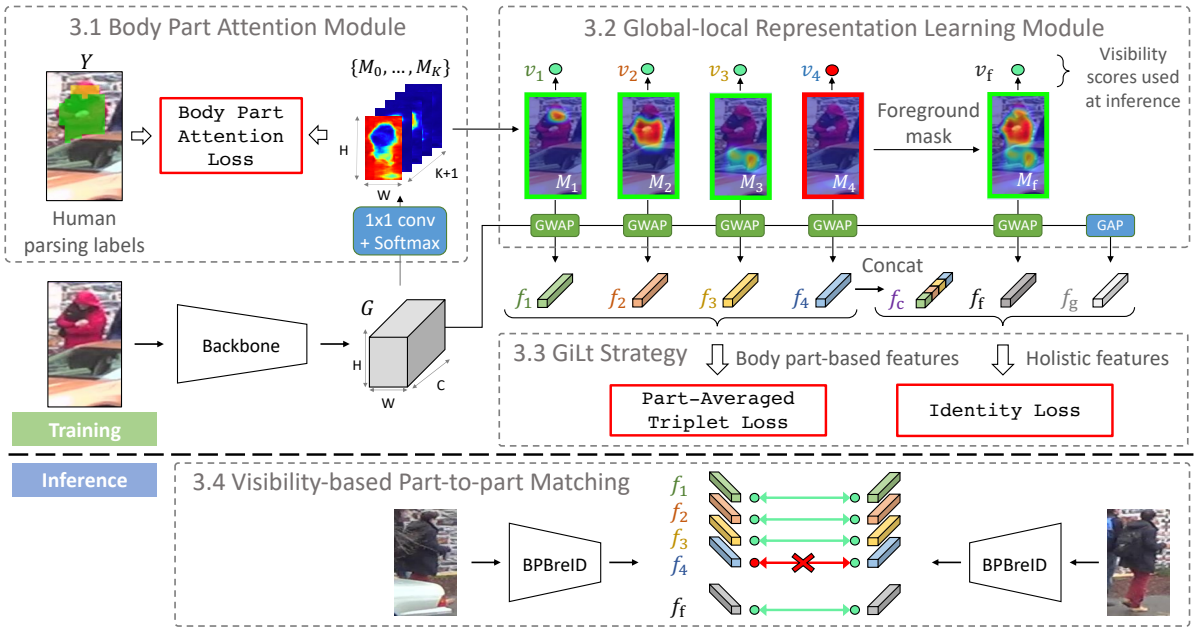


Fig. 3.2 Structure of BPBreID with detailed architecture and training procedure in the top part, and inference procedure in bottom part. The model consists of a *body part attention module* for body part attention maps and a *global-local representation learning module* for producing holistic features $\{f_g, f_f, f_c\}$ and body part-based features $\{f_1, ..., f_K\}$ together with their visibility scores $\{v_f, v_1, ..., v_K\}$. For holistic features, "g" stands for "global", "f" for "foreground" and "c" for "concatenated". *GWAP* stands for global weighted average pooling. The network is trained in an end-to-end fashion using a *body part attention loss* for supervising part prediction, a standard *identity loss* on holistic features and a *part-averaged triplet loss* on body part-based features. Query to gallery distance is computed at inference using a *part-to-part matching strategy* for comparing only mutually visible body parts. Green/red color depict visible/invisible body parts. Each component of the architecture is framed with a grey rectangle, with its name and a number referencing the section describing it. For conciseness, BPBreID is represented here with $K = 4$: {head, torso, legs, feet}.

Body Part Attention Loss

The pixel-wise part classifier is supervised with a body part attention loss L_{pa}, which is in practice a cross-entropy loss with label smoothing [68, 69], as formulated here:

$$L_{pa} = -\sum_{k=0}^{K}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} q_k \cdot log(M_k(h,w)) \,, \\ \text{with } q_k = \begin{cases} 1 - \frac{N-1}{N}\varepsilon & \text{if } Y(h,w) = k \\ \frac{\varepsilon}{N} & \text{otherwise} \,, \end{cases} \tag{3.2}$$

where the human parsing labels map Y is described in Section 3.3.1, N is the batch size, ε is the label smoothing regularization rate and $M_k(h,w)$ is the prediction probability for part k at spatial location (w,h), as described in Eq. (3.1).

3.3.2 Global-local Representation Learning Module

The global-local representation learning module takes as input the body part attention maps generated by the previous module, and outputs holistic and body part-based features of the ReID target, together with a visibility score for each part. It can be visualized in the top right part of Figure 3.2. Part-based representations, combined with their visibility scores, is our solution for achieving part-to-part matching, and solving challenges $\langle 1 \rangle$ and $\langle 2 \rangle$ from Section 3.1.

Holistic and Body Part-based Features

As described in Section 3.3.1, the body part attention module produces K spatial heatmaps highlighting the corresponding K predicted body parts of the input image.

We first combine the K body part maps $\{M_1, ..., M_K\}$ in a single foreground heatmap $M_\mathrm{f} \in R^{H \times W}$: $M_\mathrm{f}(h,w) = max\big(M_1(h,w), ..., M_K(h,w)\big)$. These heatmaps are then used to perform $K+1$ *global weighted average pooling* (denoted *GWAP* in Figure 3.2) of the appearance feature map G, to obtain the foreground embedding f_f and the K body part-based embeddings $\{f_1, ..., f_K\}$:

$$f_i = \frac{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} G(h,w) M_i(h,w)}{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} M_i(h,w)}, \; \forall \, i \in \{\mathrm{f}, 1, \ldots, K\} \,. \tag{3.3}$$

The initial global appearance feature map G is also globally average pooled (GAP) to obtain the global embedding f_g: $f_\mathrm{g} = GAP(G)$. A last embedding $f_\mathrm{c} \in R^{(C \cdot K)}$ is also produced by concatenating the K body

part-based features along the channel dimension: $f_c = concat(f_1, ..., f_K)$. Our global-local representation learning module produces therefore three holistic embeddings $\{f_g, f_f, f_c\}$ and K body part-based embeddings $\{f_1, ..., f_K\}$.

Body Part Visibility Estimation

To detect occluded body parts, we compute a binary visibility score v_i for each embedding, with $0/1$ corresponding to invisible/visible parts respectively. In our BPBreID model, visibility scores are only used at inference. For all holistic embeddings, visibility scores are set to one, i.e., $v_g = v_f = v_c = 1$. For body part-based features, visibility score v_i with $i \in \{1, ..., K\}$ is set to 1 if at least one pixel in M_i has a value above threshold λ_v, which is empirically set to 0.4, as formulated below:

$$v_i = \begin{cases} 1 & \text{if } \max_{h,w}(M_i(h,w)) > \lambda_v \\ 0 & \text{otherwise .} \end{cases} \tag{3.4}$$

3.3.3 Overall Training Procedure

The overall objective function used to optimize the network during training stage is formulated as follows:

$$L = \lambda_{pa} L_{pa} + L_{GiLt} , \tag{3.5}$$

where L_{pa} is the body part attention loss supervised with human parsing labels (introduced in Section 3.3.1) and L_{GiLt} is our *GiLt* loss, supervised with identity labels. Parameter λ_{pa} is used to control the overall part attention loss contribution and is empirically set to 0.35.

GiLt Loss

To supervise model training with the identity labels, our GiLt loss relies on two losses: the popular identity classification loss and a custom part-averaged triplet loss, which is a variant of the batch hard triplet loss [42]. However, we must carefully choose which loss to apply on each of the $K + 3$ embeddings produced by our model.

First, unlike other popular part-based method [17, 19, 21, 23, 24, 26, 56, 70], we do not apply the identity loss on part-based features because of occlusions and non-discriminative local appearance, as introduced in Section 3.1. Indeed, a part-based feature is not always discriminative enough to identify a person, which renders an identity prediction objective impossible to fulfil. Consequently, adding an identity loss on such local representation would be

destructive to performance. However, similar to most state-of-the-art ReID methods, we still benefit from the identity loss supervision by applying it on the holistic features.

Secondly, we apply the triplet loss constraint on part-based features, via our custom *part-averaged triplet loss* detailed in Section 3.3.3. At inference stage, distance between samples will be computed using these part-based features, and it therefore make sense to optimize their relative distances directly with a triplet loss constraint. However, we argue the triplet constraint should not be enforced on holistic embeddings because of occlusions. Indeed, two holistic embeddings of the same identity will have intrinsically different representations if at least one of the two is partially occluded, because each embedding will represent a different subset of the whole target body. Therefore, pulling those two holistic features close together in the feature space with a triplet loss would be destructive to performance.

In summary, we claim the best training strategy for part-based methods is to apply (i) the identity loss constraint on holistic features only and (ii) the triplet loss constraint on part-based features only, via a our custom part-averaged triplet loss. We call this strategy *Global-identity Local-triplet* or simply *GiLt*, and formulate it in our GiLt loss:

$$L_{GiLt} = L_{id} + L_{tri} = \sum_{i \in \{g,f,c\}} L_{CE}(f_i) \; + L_{tri}^{parts}(f_1, ..., f_K), \tag{3.6}$$

where L_{CE} is the cross-entropy loss with label smoothing [68] and BNNeck trick [15], and L_{tri}^{parts} is our part-averaged triplet loss detailed further below. L_{id} optimizes the network to predict the input sample identity from each holistic embedding $\{f_g, f_f, f_c\}$.

We provide extensive ablation studies in Section 3.4.4 for validating our claim. These experiments also demonstrate the superiority of our GiLt strategy for training part-based methods compared to other combination of triplet and identity losses. To our knowledge, we are the first to suggest such combination of triplet and identity losses for training part-based methods. We are also the first to conduct extensive experiments to demonstrate the impact of both losses on training performance when enforced on holistic and part-based embeddings. GiLt is illustrated in Figure 3.2.

Part-Averaged Triplet Loss

Our part-averaged triplet loss differ from the standard batch hard triplet loss [42] w.r.t. the strategy used to compute the distance between two samples. Indeed, it relies on the average of pairwise parts distances between

two samples i and j. This part-averaged distance is computed using all body part-based features $\{f_1, ..., f_K\}$ jointly:

$$d^{ij}_{parts} = \frac{\sum_{k=1}^{K} dist_{eucl}(f^i_k, f^j_k)}{K} , \tag{3.7}$$

where $dist_{eucl}$ refers to the euclidean distance. Similar to [42], the part-averaged triplet loss is then computed using the hardest positive and hardest negative part-averaged distances d^{ap}_{parts} and d^{an}_{parts} respectively:

$$L^{parts}_{tri}(f^a_0, ..., f^a_K) = [d^{ap}_{parts} - d^{an}_{parts} + \alpha]_+ , \tag{3.8}$$

where the distances from anchor sample to the hardest positive and negative samples are denoted by d^{ap} and d^{an} respectively, and α is the triplet loss margin. Therefore, our part-averaged triplet loss globally optimize an average of local distances between corresponding parts, and not a distinct triplet for each part, as adopted in [19, 23] and shown to be inferior in Table 3.2, under "*BPBreID w/o part-averaged triplet loss*". This critical design choice gives each training step the opportunity to focus on the parts with most robust and discriminant features, which in turns mitigates the impact of occluded and non-discriminative local features.

3.3.4 Visibility-based Part-to-Part Matching

Given a query sample q and a gallery sample g, pairwise distance is computed at inference by a visibility-based part-to-part matching strategy using the foreground embedding and the body part-based embeddings:

$$dist^{qg}_{total} = \frac{\sum\limits_{i \in \{f,1,...,K\}} \left(v^q_i \cdot v^g_i \cdot dist_{eucl}(f^q_i, f^g_i) \right)}{\sum\limits_{i \in \{f,1,...,K\}} \left(v^q_i \cdot v^g_i \right)} . \tag{3.9}$$

Visibility scores $v^{q|g}_i$ are used to ensure that only mutually visible body parts are compared. If there's no mutually visible part between the two samples, their distance is set to infinity. The strategy is illustrated in the bottom part of Figure 3.2. Global and concatenated embeddings are not used at inference because they may convey information from occluding objects and pedestrians.

3.4 Experiments

3.4.1 Datasets and Evaluation Metrics

We evaluate our model on the holistic dataset Market-1501 [14], and the occluded dataset Occluded-ReID[1] [29]. *Unlike the original WACV23 paper, we focus our study on two datasets, excluding results on DukeMTMC and its derivatives. This decision follows the withdrawal of the DukeMTMC-reID dataset and, by extension, related datasets such as Occluded-Duke and P-DukeMTMC. We refer readers to the original papers for more experimental results.* We report two standards ReID metrics: the cumulative matching characteristics (CMC) at Rank-1 and the mean average precision (mAP). Performances are evaluated without re-ranking [71] in a single query setting.

3.4.2 Implementation Details

Model architecture A *ResNet-50* (RN) [72] is employed as the main backbone feature extractor. The final fully connected layer and global average pooling layer are removed and the stride of the last convolutional layer is set to 1 instead of 2. For a fair comparison with methods using heavier architecture or training, we also employ the *ResNet-50-ibn* (*RI*) [73] as in [74–76], and the *HRNet-W32* (HR) [77] backbone as in [20]. HRNet feature maps with higher resolution are particularly beneficial to BPBreID for building fine-grained attention maps. All backbones are pre-trained on ImageNet [78]. The number of body parts K is set to 5 for holistic datasets and 8 for occluded datasets.

Training procedure The training procedure is mainly adopted from BoT [15]. All images are resized to 256×128 for *ResNet-50* (*RN*) and 384×128 with *HRNet-W32* (*HR*) and *ResNet-50-ibn* (*RI*). Images are first augmented with random cropping and 10 pixels padding, and then with random erasing [79] at 0.5 probability. A training batch consists of 64 samples from 16 identities with 4 images each. The model is trained in an end-to-end fashion for 120 epochs with the Adam optimizer on one NVIDIA Quadro RTX8000 GPU. The learning rate is increased linearly from 3.5×10^{-5} to 3.5×10^{-4} after 10 epochs and is decayed to 3.5×10^{-5} and 3.5×10^{-6} at 40^{th} epoch and 70^{th} epoch respectively. The label smoothing regularization rate ε is set to 0.1 and triplet loss margin α is set to 0.3.

[1] Occluded-ReID has no train set, so we use Market-1501 for training.

Table 3.1 Comparison of BPBreID with SOTA methods. Symbols † / ‡ denote *global* / *part-based* methods respectively. First, second and third best performance are indicated with [1,2,3] respectively.

Methods	Holistic		Occluded	
	Market -1501		Occluded -reID	
	R-1	mAP	R-1	mAP
BoT-based training schemes with single ResNet-50 backbone				
BoT [15] †	94.5	85.9	58.4	52.3
SGAM [66] ‡	91.4	67.3	-	-
PGFA [24] ‡	91.2	76.8	-	-
MHSA [65] ‡	94.6	84.0	-	-
VGTri [21] ‡	-	-	**81.0**	**71.0**
OAMN [80] †	93.2	79.8	-	-
HG [81] †	**95.6**	86.1	-	-
$BPBreID_{RN}$ ‡	95.1	**87.0**	76.9	68.6
Arbitrary backbones/training schemes or heavier architectures				
PVPM [17] ‡	-	-	66.8	59.5
HOReID [19] ‡	94.2	84.9	80.3	70.2
ISP [20] ‡	95.3	88.6	-	-
PAT [23] ‡	95.4	88.0	81.6	72.1
PGFL [82] ‡	95.3	87.2	80.7	70.3
HPNet [25] ‡	-	-	**87.3**[1]	77.4[3]
SSGR [76] †	**96.1**[1]	89.3	78.5	72.9
FED [83] †	95.0	86.3	86.3[2]	79.3[2]
LDS [75] †	95.8[2]	**90.3**[1]	-	-
PFD [18] ‡	95.5	89.7[2]	81.5	**83.0**[1]
$BPBreID_{RI}$ ‡	95.7	88.4	77.0	70.9
$BPBreID_{HR}$ ‡	95.7[3]	89.4[3]	82.9[3]	75.2[4]

3.4.3 Comparison with State-of-the-Art Methods

Our comparison with existing methods is constrained by the removal of the DukeMTMC-ReID dataset and its derivatives. We refer readers to the original papers for further details. We compare our model in Table 3.1 with other ReID works and it ranks first overall. Methods in the first part of the table use a ResNet-50 backbone with a training procedure similar to ours and BoT [15]. Methods in the second part either use arbitrary training procedures with bigger image size [17,25,75,82], more advanced backbones [18,20,76,83], or heavier architecture with additional backbones or branches [17–19,23,75,82].

Market-1501: Our method outperforms all part-based methods (‡) on Market-1501, except for PFD [18], which uses a much heavier architecture, with a ViT backbone and a HRNet-W48 parallel branch for pose estimation. Regarding global methods (†), BPBreID achieves competitive performance on Market-1501, although the performance difference between most SOTA methods remains insignificant on it. This demonstrates that part-based methods are a competitive choice for holistic person ReID.

Occluded-ReID: This occluded dataset requires strong domain adaptation capacity, since it does not provide a training set, and the Market-1501 dataset that is generally used for pre-training does not contain occluded samples. All well-performing methods on Occluded-reID rely on the information from an external pose estimation model at inference [17–19,82] or on occlusion data augmentation techniques [76,83] to achieve robust part pooling on the new occluded domain. Different from these methods, we don't use any external model at inference, nor occlusion data augmentation, but still achieve competitive performance.

3.4.4 Ablation Studies

In this section, we conduct some ablations studies using $BPBreID_{RN}$, to analyze the impact of our architectural choices on ReID performance.

Components of BPBreID

Performance gain related to different components of our model are reported in Table 3.2. We adopt Bag of Tricks (BoT) [15] as a baseline and build BPBreID on top of it.

BPBreID without learnable attention is an alternative method where the K body part probability maps $\{M_1, ..., M_K\}$ predicted by the body part attention module are replaced by fixed attention weights derived directly from PifPaf output. The decreased performance primarily reveals that the

Table 3.2 Ablation study for the main components of BPBreID. For the experiment "*w/o part-avgd triplet loss*", we replace our part-averaged triplet loss introduced in Eq. (3.8) by a distinct triplet loss for each part.

Methods	R-1	mAP
BoT [15] baseline	51.4	44.7
BPBreID	**66.7**	**54.1**
- w/o learnable attention	51.6	39.2
- w/o visibility scores	52.6	45.3
- w/o part-avgd triplet loss	64.8	51.7

Table 3.3 Performance comparison for body part and holistic embeddings.

Methods	R-1	mAP
f_g	60.2	47.5
f_f	64.1	49.7
f_c	64.4	50.3
f_1 (head)	47.1	24.7
f_2 (torso)	52.0	31.7
f_3 (left arm)	55.7	34.4
f_4 (right arm)	56.7	34.1
f_5 (legs)	22.3	13.3
f_6 (feet)	16.1	9.0
$\{f_1, ..., f_6\}$	65.9	52.2
$\{f_f, f_1, ..., f_6\}$	**66.1**	**52.5**

lack of end-to-end training on the attention weights leads to a discrepancy between the fixed attention masks and the ReID need in terms of backbone feature pooling. This confirms that training the attention mechanism in an end-to-end fashion with both body part prediction and ReID as objectives leads to an attention mechanism which is more specialized towards solving the ReID task, with a better selection of discriminative appearance features.

BPBreID without visibility scores refers to a strategy where all embeddings are used at inference no matter their visibility, i.e., all visibility scores $v_i^{q|g}$ in Eq. (3.9) are set to 1. As expected, using noisy embeddings corresponding to non-visible parts dramatically reduces performance. This validates the effectiveness of our visibility-based part-to-part matching strategy for solving challenges $\langle 1 \rangle$ and $\langle 2 \rangle$. Our attempts to account for the visibility scores at training did not lead to performance improvement, but remains a promising path for future research. We speculate this happens because (1) GiLt is already robust to occlusion and (2) most training samples are non occluded.

Finally, *BPBreID without part-averaged triplet loss* refers to a model where part-based embeddings are supervised individually with a classic triplet loss [42] instead of our part-averaged triplet loss described in Eq. (3.8). The difference between these two approaches lies in their underlying objective: the part-averaged triplet loss optimizes a global distance between two samples, which is computed using the average of local distances, whereas the

Table 3.4 Impact of *identity loss* and *triplet loss* on training performance when applied selectively on holistic embeddings (*global* "g", *foreground* "f" and *concatenated* "c") and *body part-based embeddings* ("$p_{1,..,K}$"). Triplet loss on $p_{1,..,K}$ refers to our part-averaged triplet loss described in Section 3.3.3. Triplet loss on other embeddings (g, f and c) refers to a standard batch-hard triplet loss [42]. Identity loss on $p_{1,..,K}$ refers to a identity loss applied individually on each part-based embeddings. We also report performance for the popular part-based ReID architecture PCB [26], which does not use a foreground embedding.

Idx	Identity loss				Triplet loss				BPBreID (ours)		PCB [26]	
	g	f	c	$p_{1,..,K}$	g	f	c	$p_{1,..,K}$	R-1	mAP	R-1	mAP
GiLt	✓	✓	✓					✓	**66.7**	**54.1**	54.6	**46.3**
PCB				✓					57.2	43.2	51.2	40.8
1				✓	✓	✓	✓		52.9	43.2	50.2	40.9
2	✓	✓	✓	✓	✓	✓	✓	✓	59.5	48.2	51.1	42.8
3	✓	✓	✓	✓					61.5	49.5	52.1	44.8
4					✓	✓	✓	✓	53.9	41.9	45.5	37.6
5	✓	✓	✓	✓				✓	61.8	49.4	51.0	43.5
6	✓	✓						✓	65.5	51.4	52.9	43.5
7	✓		✓					✓	56.5	41.9	-	-
8		✓	✓					✓	64.0	52.9	**56.2**	46.2
9	✓	✓	✓				✓	✓	66.2	53.3	55.9	46.1
10	✓	✓	✓			✓		✓	63.6	52.2	-	-
11	✓	✓	✓		✓			✓	64.0	52.4	54.8	45.9
12	✓	✓	✓				✓		65.3	52.9	54.4	45.7

standard triplet loss optimizes all local pairwise distances individually. The later approach gives reduced performance because it renders the training procedure more sensitive to occluded body-parts and non-discriminative local appearance. This last ablation test demonstrates the robustness of our part-averaged triplet loss to occlusions and non-discriminative local features, and therefore to solve challenges $\langle 1 \rangle$ and $\langle 3 \rangle$.

Validation of the GiLt Strategy

In this section, we study the impact of different combinations of the identity and triplet losses on the $K + 3$ embeddings produced by our model. Results are reported in Table 3.4. We also report performance for the popular model architecture PCB [26], which partition the input image in six horizontal

stripes, to demonstrate the superiority of our training scheme with other part-based architecture. The original PCB paper suggest a simple identity loss applied on part-based embeddings only: the corresponding sub-optimal performance is reported in the second table row. The experiment on the first row correspond to our *GiLt* strategy described in Section 3.3.3: holistic features are supervised only with an identity loss and part-based features are supervised with our part-averaged triplet loss. As demonstrated by experiments 1 to 4, triplet and identity losses are complementary to each other and best performance is reached when using them together. However, naively applying both losses on all embeddings (experiment 2) is a sub-optimal solution. We can draw two conclusions from experiments 5 to 8, which are small variations of our GiLt strategy regarding the identity loss. First, applying the identity loss on all three holistic embeddings leads to a more robust training scheme and to better performance. This experiment validates our choice of computing a global and a concatenated embeddings for training, even though we don't use them at inference. Second, experiment 5 validates our intuition that using an identity loss on part-based features is harmful to performance, since it renders the training procedure sensitive to occlusions and to non-discriminative local features. Experiments 9 to 12 validate our *GiLt* strategy of enforcing the triplet loss constraint on part-based embeddings only.

Discriminative Ability of Output Embeddings

In this Section, we study the discriminative ability of the holistic and body part-based embeddings $\{f_g, f_f, f_c, f_1, ..., f_K\}$ for $K = 6$. For that purpose, we compute query-to-gallery samples distance using each embedding individually or combination of them, and report the corresponding ranking performance in Table 3.3. When multiple embeddings are used, we compute the average distance weighted by visibility scores, as described in Eq. (3.9). As demonstrated in Table 3.3, performance using holistic embeddings are sub-optimal because the global embedding is sensitive to background clutter, and the foreground embedding cannot achieve part-to-part matching. The concatenated embedding fixes those issues but remains sensitive to occlusions because it contains noisy information from embeddings of non visible body parts. As demonstrated in the table, using body part-based embeddings individually leads to sub-optimal performance. Performance is better for embeddings from upper body parts, because (1) these parts are more discriminative and (2) lower body parts are very often occluded. Using all parts embeddings $\{f_1, ..., f_6\}$ leads to the best performance, because

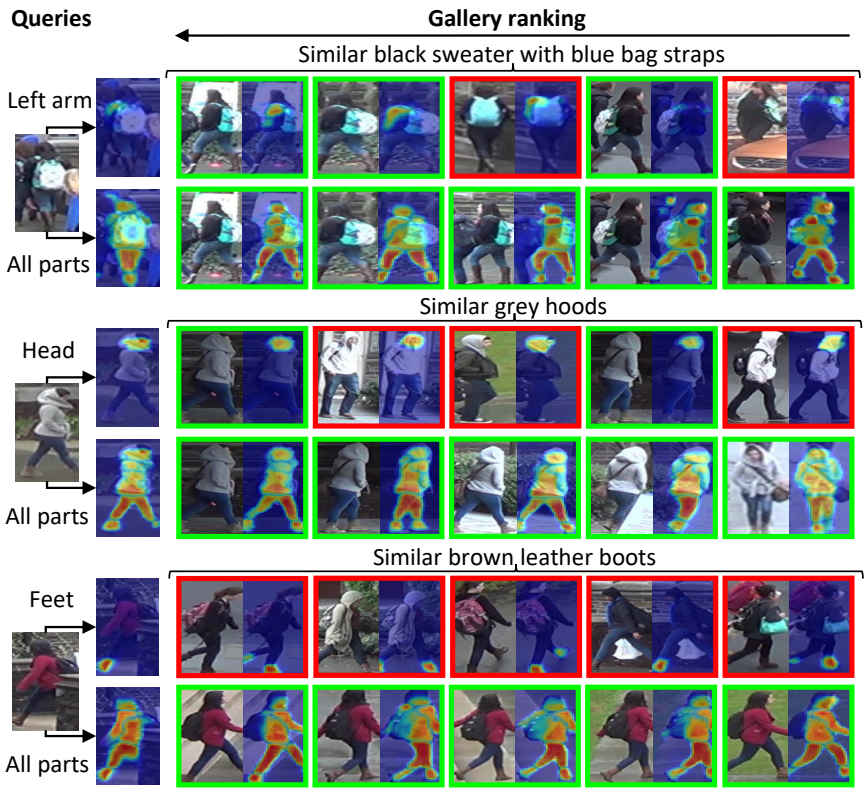


Fig. 3.3 Visualization of ranking results based on individual body part-based embeddings (top row of each query) or all body part-based embeddings with the foreground embedding (bottom row of each query). For the "all parts" rows, only the foreground attention map is displayed for conciseness. In the top row of each query, the retrieved gallery samples are very similar w.r.t. the compared body part, but identities do not match because a single body part is not discriminative enough. Green/red borders are correct/incorrect matches. Best viewed in color and zoomed in.

this strategy is the key to overcome the big challenges related to occluded re-id, i.e., (1) achieve feature alignment, (2) reduce background clutter and (3) compare only mutually visible body-parts. Finally, adding information from the foreground embedding produces slightly better performance, because it helps in mitigating errors caused by failed body part prediction, and by image pairs having few or no mutually visible parts. Figure 3.3 illustrates some ranking results using these embeddings individually.

3.5 Conclusions

In this work, we propose our model BPBreID to address the occluded person ReID task by learning body part representations and make two contributions. First, we design a body part attention module trained from a dual supervision with both identity and human parsing labels. With this attention mechanism, we show how external human semantic information can be effectively leveraged to produce ReID-relevant part-based features. Second, we investigate the influence of triplet and identity losses for learning part-based features and provide a simple yet effective GiLt strategy for training any part-based method. Our model achieves state-of-the-art performance on five popular ReID datasets.

4

KPR: Keypoint Promptable Re-Identification

Published at ECCV 2024

4.1 Introduction

Person re-identification (ReID) [12] is a challenging retrieval task that involves matching a query image of a person of interest with other person images. ReID finds wide-ranging applications in multi-object tracking [84], pedestrian flow analysis [85], sport understanding [4–6, 8, 9] and video-surveillance [14]. However, ReID is a difficult task due to various factors including inaccurate bounding boxes, luminosity changes, poor image quality, and occlusions.

In recent years, there has been a growing interest in addressing the occluded re-identification task [13], where the ReID target might be occluded by various objects or other people. In this work, we identify and explicitly address a critical challenge that has been overlooked by previous occluded ReID methods [1,83]: the presence of **Multi-Person Ambiguity** (MPA) inherent to bounding box annotations. The multi-person ambiguity arises when a bounding box image contains multiple individuals, making it challenging even for humans to accurately identify the intended ReID target among all the candidates. Figure 4.2a shows how MPA can negatively impact the re-identification model, causing feature mixing among different individuals or even a focus on an unintended person.

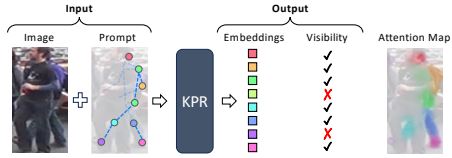
To tackle this multi-person ambiguity, it is necessary to incorporate additional pixel-level information, such as keypoints or segmentation masks, provided by an upstream human operator or vision model. This additional data plays a vital role in assisting the downstream ReID method to distinguish the intended target from other candidates.

Inspired by recent advances in promptable[1] methods for vision transformers (e.g., SAM [86]), we propose to explicitly use additional semantic[2] keypoints as inputs to disambiguate images with multiple individuals. Semantic keypoints are suitable for two crucial use cases: 1) manual prompting by a human operator with a few clicks on an image, and 2) automated prompting by an upstream pose estimation model. Nevertheless, our architecture could be easily extended to support segmentation masks as prompts.

Prior research [17–19] on *pose-guided methods* has explored the use of additional information from external pose estimation models. However,

[1] A "prompt" typically refers to an input instruction that guides the model's response. In vision tasks, e.g. segmentation, prompts can be any kind of data that specifies what the model should focus on [86].

[2] Semantic keypoints are distinct points within a person image that carry semantic information about specific body parts, typically obtained with pose estimation [87].



(a) Keypoint Promptable ReID (KPR)

(b) Part-to-part matching

(c) Back (occluded) person prompted

(d) Front person prompted

Fig. 4.1 Overview of our proposed Keypoint Promptable ReID (KPR) method. KPR takes an image with keypoints prompts as input and produces part-based features of the prompted target. The prompt instructs the model to focus on a specific individual, i.e. the back blue jacket man (b) or the front black t-shirt man (c) in this example. Colored dots illustrate the positive keypoints prompt, with one color per body part.

none of these works explicitly tackled the multi-pedestrian ambiguity issue.

Following above insights, we propose **KPR**, a novel ***K**eypoint-**P**romptable transformer for part-based person **R**e-identification*, illustrated in Fig. 4.1a. KPR is a transformer-based model that takes an image along with *semantic keypoints* as input. It then produces *part-based features*, each representing a distinct body part of the ReID target, along with their respective visibility scores. Part-based methods have demonstrated superior performance [1,20] in re-identifying occluded individuals because they utilize only the visible parts for comparison. Our method can process both *positive* and *negative* keypoints, which respectively represent the target and non-target pedestrians. As illustrated in Fig. 4.1c, prompts can be used to re-identify the desired (front or back) person in case of occlusion. Furthermore, KPR is designed to be *prompt-optional* to offer more practical flexibility. This means the same model can be used without prompt on non-ambiguous images, or with prompt when dealing with occlusions, while consistently achieving state-of-the-art performance in both cases. To demonstrate the advantages of keypoint prompts for re-identification, we evaluate KPR on two tasks: *person retrieval* and *multi-person pose tracking*.

(a) Existing ReID models (e.g. BPBreID [1]) are confused by **MPA**

(b) MPA explicitely addressed with positive and negative keypoints prompt

Fig. 4.2 Person retrieval with Multi-Person Ambiguity (**MPA**). Green/red borders denoted correct/incorrect matches. Red/pastel dots indicate negative/positive prompts.

Moreover, since no current ReID dataset is provided with explicit target identification in case of multi-person ambiguity, we introduce a new dataset called ***Occluded PoseTrack-ReID***, nicknamed "*Occ-PTrack*", and derived from the PoseTrack21 dataset [88]. This dataset offers high-quality keypoints annotations and features a substantial number of images with multi-person occlusions.

Overall, we summarize the main contributions of our work as follows:

1. We propose *KPR*, a novel prompt-optional model for part-based ReID. To our knowledge, we are the first to introduce a visual prompting mechanism for person re-identification, and the first to explicitly address the ambiguity caused by multi-person occlusions.
2. We introduce *Occ-PTrack*: the first multi-person occluded ReID dataset with explicit target identification through manual keypoints annotations. Furthermore, we propose new keypoint annotations for four popular re-identification datasets.
3. Our method outperforms all previous state-of-the-art methods on the occluded ReID task, demonstrating the effectiveness of the keypoint promptable approach.

Our ambition is to advocate for a shift from the ambiguous bounding-box approach to ReID and promote a keypoint-centric paradigm. To this end, we publicly release our codebase, proposed dataset, and annotations, to provide a common setup for evaluating promptable ReID methods.

4.2 Related Work

Multi-Person Ambiguity: A previous study [83] addressed a related issue called "Non-Target Pedestrians (NTP)," similar to our concept of "Multi-

Person Ambiguity" (MPA). However, their focus was on mitigating the negative impact of NTP during training. In contrast, we provide a method to explicitly disambiguate the intended ReID target during both training and testing phases.

Pose-guided ReID: We detail here the fundamental difference between our novel keypoint promptable approach and the traditional pose-guided methods. Although none of the previous works explicitly address the multi-person ambiguity (MPA) issue, we identify two categories within these pose-guided approaches: 1) those inherently unable to address MPA [1, 17, 21, 22, 24, 76, 80, 89], and 2) those with the potential to overcome MPA [18, 19]. In the first category, methods like BPBreID [1] or Pirt [22] utilize pose labels only during training, and can not leverage additional information at test time to disambiguate multiple persons. Moreover, methods relying on horizontal stripes (VGTri [21], PGFA [24]) or the affinity fields of a pose model (PVPM [17]), are meant to tackle object-occlusions, but struggle to disambiguate multiple persons. In the second category, methods like HOReID [19] and PFD [18] directly employ keypoints from a single person, enabling focus on a specified target. However, HOReID [19] is restricted to utilizing spatial features at the exact keypoints locations, thus overlooking critical appearance cues on entire body regions. Finally, PFD [18], and all previous pose-guided methods including HOReID, incorporate pose information *after appearance encoding has occurred*, e.g., for local pooling of a spatial feature map, but do not leverage pose data to guide the appearance encoding process as our method do. Indeed, our approach implements a *prompt-aware appearance encoding*, since it conditions feature extraction on the input keypoints. This enables *better feature disentanglement between the target and the occluders* from the early feature extraction stages, resulting in more accurate representations. Furthermore, our method is *prompt-optional*, setting it apart for its flexibility from these previous pose-guided methods. Finally, our method can ingest additional *negative keypoints*, to further disambiguate a multi-person occlusion scenario.

Prompting in Vision: Recent works have introduced promptable transformers to tackle object segmentation. For instance, SAM [86] employs two networks: a transformer encoder to generate generic spatial features and a lightweight promptable decoder to integrate user inputs for interactive segmentation. In contrast, our approach directly prompts a feature encoder, so that *feature extraction is explicitly driven by the prompts*. Unlike SimpleClick [90], which prompts a transformer with a single 2D point and a previous mask for iterative segmentation, our model ingests multiple key-

points concurrently, enhancing body part localization and representation by *leveraging their semantic information*.

4.3 Methodology

The overall architecture of our model KPR is illustrated in Fig. 4.3 and is based on a Swin transformer backbone [91]. The next subsections first describe the various components of the architecture. We then present the training process, including our proposed data augmentation. Finally, we describe the inference procedure and conclude with an intuitive discussion of our design choices.

4.3.1 Tokenization

The tokenization procedure coverts raw image and keypoints information into C_i-dimensional tokens, to be subsequently processed by the Swin encoder.

Image tokenization: The input image, with dimensions $H \times W$, undergoes tokenization using the conventional Swin patch embed module. This process generates $H_t \times W_t = \frac{H}{4} \times \frac{W}{4}$ spatial tokens of dimensions C_i by embedding each 4×4 image patch through a linear layer.

Keypoints tokenization: As previously mentioned, our model uses two types of semantic keypoints as prompts: *positive keypoints* indicating the ReID target and *negative keypoints* indicating other individuals. We organize these prompt keypoints into K+1 groups: the first group contains all negative keypoints, and the remaining K groups contain subsets of the positive keypoints. Each subset of positive keypoints actually corresponds to a semantic body region. For instance, with $K = 8$, we define groups like: {*head*, *torso*, *right/left arm*, *right/left leg*, and *right/left feet*}. For each group $i \in \{1, ..., K+1\}$, a unique heatmap M_i of size $H \times W$ is created by drawing a Gaussian 2D kernel at each of its keypoint's center location (x, y). The kernel's standard deviation is set to a fraction α_G of the image width. Invisible keypoints are ignored. These $K+1$ heatmaps are then concatenated into a global spatial tensor M of size $H \times W \times (K+1)$. Finally, each 4×4 patch in M is linearly projected to produce prompt tokens of dimension C_i. Like the image tokenizer, the keypoints tokenizer outputs therefore $H_t \times W_t$ spatial tokens of dimension C_i.

Token fusion: Images and keypoints tokens are fused by summing tokens at the same spatial location, producing $H_t \times W_t$ spatial tokens with

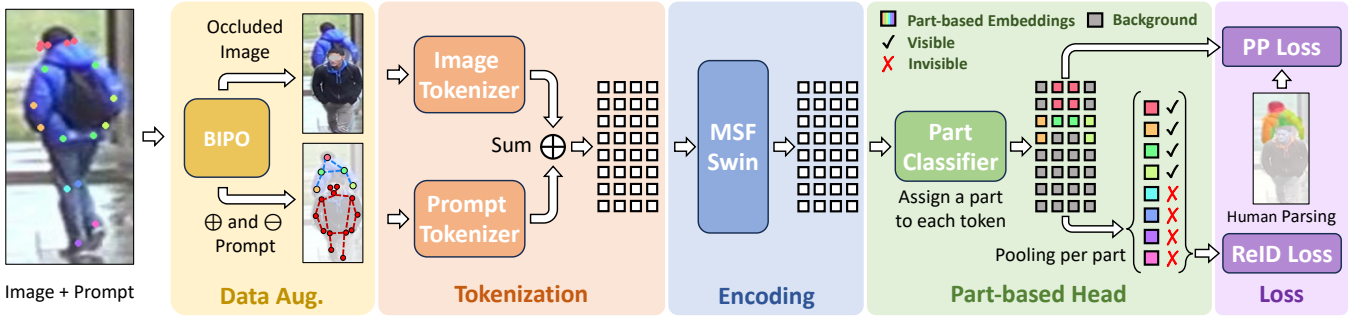


Fig. 4.3 Architecture overview of our proposed Keypoint Promptable ReID (KPR) model. The Batch-wise Inter-Person Occlusion (BIPO) data augmentation is first applied to generate artificial inter-person occlusions (Sec. 4.3.5). The image and the optional positive/negative ($\oplus/\ominus$) prompts are then tokenized and summed (Sec. 4.3.1). Tokens are then fed to our proposed Multi-Stage feature Fusion (MSF) Swin transformer to generate high-resolution feature maps (4.3.2). The feature map is then fed to a Part-based Head (PBH), which assigns a part (or the background) to each token with a part classifier, and then averages all tokens of the same part to produce the final K part-based embeddings $\{f_1, ..., f_K\}$ and their binary visibility scores $\{v_1, ..., v_K\}$, with $v_i \in \{0, 1\}$, and visually denoted by {✗, ✓} (Sec. 4.3.3). A part i with no token assigned is considered invisible (i.e., $v_i = 0$), and ignored when computing two samples' similarity. KPR is illustrated here for K=8, with a unique color for each body part: {*head*, *torso*, *right/left arm*, *right/left leg*, and *right/left feet*}. Finally, the entire pipeline is trained with two losses: a Part-Prediction (PP) Loss and a ReID Loss (Sec. 4.3.4).

dimension C_i. These tokens are then forwarded to the Swin-based encoder.

Design motivations: Prompted keypoints are organized into $K+1$ channels to preserve semantic meaning before tokenization. It allows the model to differentiate positive and negative prompts, and to distinguish prompt information of different body parts, for better part localization. *Prompt-aware appearance encoding*: Moreover, feeding the network with keypoint information before the encoding stage is critical in enabling the network to disentangle the ReID target features from other persons' features. Finally, this design offers the flexibility to support segmentation mask prompts by replacing M with a double-channel map of positive and negative segmentation masks. Regarding token fusion, the sum operation maintains the input dimension for the Swin backbone, permitting the use of pre-trained weights from the image-only model. *Prompt optionality*: This design also offers the flexibility of running the model without prompts, by processing the image tokens alone, with no architectural adjustments required.

4.3.2 MSF Swin Encoder

A *Swin-transformer* [91] is employed as the backbone feature extractor for its great performance in vision tasks and its ability to capture high-frequency details, which is beneficial to ReID [92]. Moreover, we propose to enhance Swin with a Multi-Stage feature Fusion strategy, denoted as "MSF", to output feature maps with a high spatial resolution. Indeed, similar to CNN, each Swin stage reduces the number of tokens by a factor of 4 while doubling the channel dimension. To obtain a high-resolution token map, the outputs of all stages are bilinearly upsampled to the same resolution $H_t \times W_t$ as the output of the tokenization module, then concatenated into a single feature map along the channel dimension, and finally passed through a linear layer to obtain the final spatial token map of size $H_t \times W_t \times C_o$. Different from the standard Swin, our MSF Swin therefore preserves the input tokens resolution. As pointed out in prior works [15, 20], a high-resolution feature map enriches the granularity of the features and brings significant performance improvements (see Tab. 4.3, Exp. 3).

4.3.3 Part-based Head

The encoder yields $H_t \times W_t$ tokens, which are sent to the Part-based Head (PBH) module, as illustrated in Fig. 4.3. This module is derived from the Global-local Representation Learning Module introduced by BPBreID [1]. Each token is first labeled either as one of the K body parts or as background,

by a *token-wise part classifier*. The part classifier is implemented as a linear layer with $K+1$ output logits followed by the softmax operation, and is applied individually on each token. The assignment of each token to one of the K body parts (or the background) is therefore formulated as a classification task with $K+1$ classes. We term the probability map of size $H_t \times W_t \times (K+1)$ produced by this part classifier over all tokens as "*part-attention maps*". These maps are showcased as colored heatmaps in Fig. 4.1a and Fig. 4.6, with one color per channel (the background channel is ignored). Finally, tokens having the same assigned body part are averaged together, resulting in K part-based embeddings $\{f_1, ..., f_K\}$ of size C_o. Background tokens are ignored. Parts with no tokens assigned are considered invisible and given a visibility score $v_i = 0$. The corresponding invisible embedding f_i, with $v_i = 0$, is therefore ignored for downstream computation. The PBH module thus produces K part-based embeddings $\{f_1, ..., f_K\}$ and their binary visibility scores $\{v_1, ..., v_K\}$.

Human parsing labels: To train the token-wise part classifier, each spatial token is assigned a ground-truth class, i.e., part label, derived from the coarse human parsing labels provided by [1]. These human parsing labels are illustrated on the right of Fig. 4.3 and in Fig. 4.4a.

4.3.4 Training Process

The overall objective function used to train the model is:

$$L = L_{\text{ReID}} + \lambda_{pp} L_{pp} \,, \tag{4.1}$$

where L_{ReID} is the ReID loss supervised with identity labels, L_{pp} is the token-wise part-prediction loss, and λ_{pp} is a weight parameter empirically set to 0.3.

Part Prediction Loss: The token-wise part classifier introduced in Sec. 4.3.3 is supervised with a Part Prediction (PP) Loss that is actually a cross-entropy loss applied on each spatial token as in [1].

Re-identification loss: As a ReID objective, we employ the GiLt Loss [1], that is specifically designed to train part-based ReID models. It has the benefit of being robust to occlusions and non-discriminative body parts. GiLt combines an identity loss [15] and a batch-hard triplet loss [42]. The distance function employed for the triplet loss is the average of all local part-based distances.

4.3.5 BIPO Data Augmentation

Since most ReID datasets feature limited occlusion in their training set, a model often struggles to handle the occlusions it encounters during testing. To address this, we propose a novel Batch-wise Inter-Person Occlusion data augmentation (BIPO)[3], to generate artificial inter-person occlusions on training images, prompts, and human parsing labels. When provided with a training sample, BIPO randomly selects a different person's image (the occluder) from the same training batch, contours it with a segmentation mask derived from the human parsing labels, and overlays it on the main image. The human parsing label and keypoints prompt of the training image are then updated accordingly. Finally, all positive keypoints from the occluder are added to the training prompt to serve as additional negative points.

Design motivations: Given that a triplet loss with batch-hard mining [42] is employed during training, using images from the same training batch to generate occluders yields hard positive and negative samples for mining triplets, thereby creating challenging scenarios for contrastive learning. BIPO is illustrated in Fig. 4.4a and its significance for KPR is further discussed in Sec. 4.3.7.

4.3.6 Inference

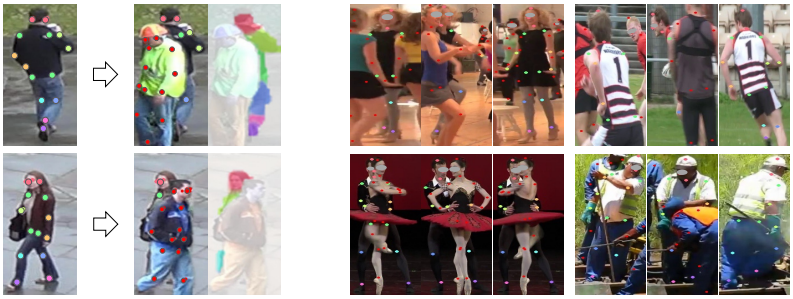
At inference, KPR processes each input image together with their respective keypoints prompts, and produces K part-based representations $\{f_1, ..., f_K\}$ of the ReID target and their binary visibility scores $\{v_1, ..., v_K\}$. Similar to previous part-based works [20, 21], the global distance $dist^{qg}_{\text{global}}$ between a query image q and a gallery image g is computed as the average cosine distance $dist_{\cos}$ of the part-based embeddings f_i that are visible in both samples:

$$dist^{qg}_{\text{global}} = \frac{\sum\limits_{i \in \{1,...,K\}} \left(v_i^q \cdot v_i^g \cdot dist_{\cos}(f_i^q, f_i^g) \right)}{\sum\limits_{i \in \{1,...,K\}} \left(v_i^q \cdot v_i^g \right)} . \tag{4.2}$$

4.3.7 KPR Pipeline Motivation and Intuition

Our pipeline comprises two closely related architectural components. First, the prompt tokenizer, which encodes semantic keypoints information to dis-

[3] Inspired by the copy-paste technique for segmentation [93].



(a) BIPO **(b)** Occluded PoseTrack-ReID

Fig. 4.4 (a) Our proposed Batch-wise Inter-Person Occlusion (BIPO) data augmentation creates artificial person occlusions that are consistent across image, prompt, and human parsing labels. BIPO is crucial to enforce the model to rely on the input prompts. (b) Four identities with their keypoints labels from our Occ-PTrack dataset.

ambiguate the ReID target, thereby providing prior information about body part location to drive feature extraction. Second, the Part-based Head (PBH), responsible for localizing the body parts of the target and constructing its part-based features.

While the Part-based Head is explicitly supervised with human parsing labels at training to learn human topology, there is no specific loss incentivizing the model to leverage information from the input prompt. Additionally, multi-person occlusions are rare in most ReID training sets, giving the model little incentive to rely on these prompts to localize the target person. For this reason, just adding a prompt tokenizer and training the pipeline as is produces a KPR model with weak reliance on the input prompts, and therefore low performance in multi-person occlusion scenarios.

To address these issues, our proposed BIPO data augmentation plays a critical role. By generating artificial multi-person occlusions during training, the model is compelled to rely more on the prompts, as they represent the sole reliable source of information to distinguish the true ReID target from the occluder.

Consequently, incorporating BIPO during training is crucial to produce a model that is proficient at leveraging prompt information and effectively handling multi-person occlusions. Finally, designing the Part-based Head without explicit reliance on the prompt offers an additional advantage: it renders our model *prompt-optional*, enhancing robustness against noisy or

missing prompts while still enabling effective localization of the target's body parts.

4.4 Experiments

4.4.1 Evaluation Benchmarks and Metrics

We assess our model's performance on three types of *person retrieval* setups: non-occluded (i.e., holistic) on **Market-1501** [14], object occlusions on **Occluded-ReID** [29] and **Partial-ReID**[4] [28], and finally multi-person occlusions on our below introduced **Occluded PoseTrack-ReID** dataset. **Occluded-Duke** [24] was taken down and removed from our study; check out our GitHub for more information. Two standard person retrieval metrics are reported: the CMC at *Rank-1* and the *mAP*. Performances are evaluated in a single query setting and without re-ranking [71]. Furthermore, we evaluate our model on *multi-person pose tracking* with the **PoseTrack21** [88] dataset. We report the *HOTA* [94] performance metric, and relevant sub-metrics such as association accuracy (*AssA*), detection accuracy (*DetA*), and the number of identity switches (*IDs*) [95].

4.4.2 Proposed Occluded PoseTrack-ReID Dataset and Annotations

We introduce Occluded PoseTrack-ReID (or simply Occ-PTrack), a new ReID dataset we built out of the annotation available with PoseTrack21 [88], a popular video benchmark for multi-person pose tracking, that features keypoints and cross-video identity annotations. Unlike previous ReID datasets focused on street surveillance, Occ-PTrack consists of images from everyday life videos, primarily from sports activities, as illustrated in Fig. 4.4b. Occ-PTrack is divided into a train/test that includes 1000/1411 identities with 17.898/13.412 images from 474/170 videos, which is roughly equivalent in terms of scale to other popular ReID datasets [14, 24]. To assess the ReID model's performance in multi-person occlusion scenarios, we select the most cluttered images of each identity in the test set as query samples, and the remaining test images as gallery samples. Cluttered images corresponds to multi-persons occlusions scenarios where either the front (occluding) or back (occluded) person is the ReID target, to evaluate the model ability to re-identify individuals in both scenarios. Occ-PTrack is challenging as

[4] We use the occluded (non-cropped) images as queries.

persons within the same video exhibit a high degree of visual resemblance since they often wear similar sports kits.

Keypoint Annotations for Standard ReID Datasets: Since conventional ReID datasets [14, 24, 29] do not contain keypoint annotations, we generate pseudo-labels for them with the PifPaf [67] pose estimator. These pseudo-labels include keypoint annotations in COCO [87] format for each image, as illustrated in Fig. 4.4a. When multiple skeletons are detected in an image, we make the assumption the one with its head closer to the top center part of the image is the intended target, and mark it with a "*is_target*" attribute. Other skeletons are used as negatives. A more detailed description of the annotation procedure can be found in the supplementary materials of the original paper.

4.4.3 Implementation Details

We evaluate our model with both the ImageNet [78] (**KPR$_{\mathbf{IN}}$**) and the SOLIDER [96] (**KPR$_{\mathbf{SOL}}$**) pre-trained Swin backbones. SOLIDER is a human-centric foundation model trained on the large-scale LUPerson [97] dataset in a self-supervised way. Following [1], the number of body parts K is empirically set to 8 for Occ-PTrack, and 5 otherwise. The training procedure is mainly adopted from TransReID [98] and SOLIDER [96]. We apply our *BIPO* data augmentation (Sec. 4.3.5) with a 0.3 probability. The model is trained for 120 epochs with a batch size of 64 and a cosine annealing *lr* scheduler. For KPR$_{\text{IN}}$/KPR$_{\text{SOL}}$, images are resized to a width of 128 and a height of 256/384. The Gaussian kernel standard deviation α_G (Sec. 4.3.1) is set to 0.1.

4.4.4 Comparison with State-of-the-Art Methods

In Table 4.1, we present a comparative analysis of KPR, when **using prompts** (*KPR*) and **not using prompts**[5] (*KPR w/o prompt*), against all leading occluded ReID methods. Our method emerges as the top performer overall.

Occluded PoseTrack-ReID: Unlike the other three datasets, Occ-PTrack contain a substantial number of images with multi-person occlusions, making them suitable environments to assess the effectiveness of our promptable method. On this challenging occluded dataset, our proposed model KPR outperforms all previous methods. Our *keypoint promptable* approach offers several key advantages over the previous *pose-guided* methods, including:

[5]The prompt tokenizer is complitely removed at both training and inference.

Table 4.1 Comparison of KPR with SOTA methods. Results in *Italic* are not provided in the original paper but reproduced by ourselves. The $1^{st}/2^{nd}/3^{rd}$ best scores are indicated with $^{1/2/3}$.

Datasets	Market-1501		Occluded-reID		Partial-reID		Occluded-PoseTrack	
Type	Holistic		Occluded					
Object Occlusions			✓		✓			
Person Occlusions							✓	
Methods	R-1	mAP	R-1	mAP	R-1	R-3	R-1	mAP
BoT [15]	94.5	85.9	58.4	52.3	-	-	*78.8*	*69.7*
PCB [26]	93.8	81.6	-	-	-	-	*81.7*	*71.2*
VGTri [21]	-	-	81.0	71.0	85.7	93.7[3]	-	-
PVPM [17]	-	-	66.8	59.5	78.3	-	-	-
HOReID [19]	94.2	84.9	80.3	70.2	85.3	91.0	-	-
ISP [20]	95.3	88.6	-	-	-	-	-	-
PAT [23]	95.4	88.0	81.6	72.1	88.0[3]	92.3	-	-
TRANS [98]	95.2	88.9	-	-	-	-	*83.5*	*73.4*
SOLIDER [96]	**96.9**[1]	**93.9**[1]	-	-	-	-	*84.4*	*76.1*[3]
SSGR [76]	96.1	89.3	78.5	72.9	-	-	-	-
FED [83]	95.0	86.3	**86.3**[1]	79.3[3]	84.6	-	-	-
BPBreid [1]	95.7	89.4	82.9	75.2	-	-	*84.9*	*75.5*
PFD [18]	95.5	89.7	83.0	81.5[2]	-	-	-	-
KPR_{IN} w/o prompt	95.6	88.7	83.3	78.2	81.7	86.0	85.3	75.4
KPR_{IN}	95.9	89.6	85.4[2]	79.1	86.0	90.0	**92.3**[1]	**82.3**[1]
KPR_{SOL} w/o prompt	96.6[3]	93.0[3]	80.0	78.5	90.3[2]	93.7[2]	86.1[3]	75.8
KPR_{SOL}	96.6[2]	93.2[2]	84.8[3]	**82.6**[1]	**90.7**[1]	**94.0**[1]	90.6[2]	81.2[2]

1) the ability to handle multi-person occlusions; 2) the capability to process both positive and negative keypoints; 3) a *prompt-aware appearance encoding*, leading to better feature disentanglement between multiple persons; and 4) the flexibility of prompts being optional.

We refer readers to Sec. 4.2 for a detailed discussion about these key differences and a comparison to previous methods. On our proposed Occ-PTrack, we can see that the best-performing method is the foundation model SOLIDER, but it still lies about 6% behind our KPR solution. Even BPBreID [1], which is specialized for occluded scenarios, demonstrates far lower performance, since it is not designed to address the Multi-Person Ambiguity. Our experiments show that integrating keypoint prompts boosts performance by at least 7.0% Rank-1 and 6.2% mAP, since it helps in mitigating the negative impact of MPA.

Market-1501, Occluded-ReID and Partial-reID: These three datasets primarily feature single individuals in their images, which means multi-

person ambiguities are rare, and therefore prompts are not exploited to their full potential. On Market-1501, our method performs on par with the state-of-the-art. However, performance levels have plateaued in recent years on this dataset. In contrast, on Occluded-ReID and Partial-reID, using a prompt noticeably improves performance by helping distinguish the individual features from occluding objects. Additionally, using prompt is beneficial when facing this cross-domain[6] setting. These findings are consistent with earlier studies [17–19,83] that also leveraged pose estimation to attain strong performance on this dataset.

4.4.5 Multi-Person Pose Tracking in Videos

We introduce a simple yet robust pose tracker, referred to as *KPRTrack*, built upon our promptable ReID method. KPRTrack combines the YOLOX [99] object detector with the HRNet [77] pose estimator. For each detected person, it extracts part-based ReID features utilizing their keypoints as positive prompts and keypoints from other individuals *within the detected person bbox* as negative prompts. Detections are matched in an online fashion with the Hungarian algorithm based on the distance function introduced in Eq. (4.2). Only tracklet-detection pairs with a distance below a threshold of 0.2 are considered for matching. Therefore, our proposed tracker relies solely on appearance information, meaning spatio-temporal cues are omitted. Pose tracking results are reported in Tab. 4.2. Despite relying on ReID only, KPRTrack achieves state-of-the-art performance, demonstrating that an occlusion-robust ReID model has strong tracking capabilities. Finally, our two last experiments in Tab. 4.2 reveal that employing keypoint prompts significantly enhances tracking performance. This improvement can be attributed to the enhanced communication between the detector and the ReID model via the keypoint prompts. This leads to improved association accuracy and reduced identity switches, especially when multiple targets intersect paths (i.e., multi-person occlusion scenarios).

4.4.6 Ablation Studies

In this section, we analyze the performance impact of our architectural choices.

Fig. 4.5 Qualitative results on multi-person pose tracking: all track IDs, depicted by unique skeleton colors, are maintained by KPRTrack after a long occlusion by the front cyclist. Colored heatmaps illustrates the part-attention maps of KPR (see Sec. 4.3.3), with one color per body part.

Table 4.2 Multi-person pose tracking performance in videos on PoseTrack21 [88].

Method	HOTA↑	DetA↑	AssA↑	IDs↓
TRMOT [100]	46.8	40.9	55.0	-
FairMOT [101]	53.5	47.4	61.4	-
Tracktor++ [52]	58.3	52v.7	65.4	-
CorrTrack + ReID [88]	56.9	51.3	64.2	-
KPRTrack w/o prompt	61.5	58.4	65.8	3830
KPRTrack	**63.0** (+1.5)	**59.0** (+0.6)	**68.3** (+2.5)	**3344** (-486)

Components of KPR

As demonstrated in Table 4.3, all our proposed modules introduced in Sec. 4.3 play a crucial role in improving the overall ReID performance. Exp. 2 and Exp. 8 demonstrate the importance of the Part-based Head to compare only mutually visible body parts when facing occluded persons, compared to using a single global feature (as in Exp. 1 and 7). PBH degrades performance when applied alone (Exp. 2 vs Exp. 1), but is beneficial when jointly enabled with other modules (Exp. 8 vs Exp 7). Exp. 3 is consistent with claims from previous works [15, 20] that increasing the feature map resolution is crucial to achieve fine-grained re-identification and better overall performance. Exp. 4 and 5 illustrate the benefits of positive and negative prompts to disambiguate the intended ReID target from other persons. Exp. 6 demonstrates the importance of our BIPO augmentation to generate artificial occlusions at training, since most occluded ReID benchmarks feature

[6]Occ-ReID and Partial-ReID have no train set, so Market-1501 is used for training.

Table 4.3 Ablation study of our proposed KPR architecture. PBH stands for "Part-based Head", meaning we use multiple part-based embeddings with visibility scores instead of a single global embedding (Sec. 4.3.3). MSF stands for the "Multi-Scale Features" module added to Swin (Sec. 4.3.2). Prompt $\oplus/\ominus$ refers to positive/negative prompts (Sec. 4.3.1). BIPO stands for the "Batch-wise Inter-Person Occlusion" data augmentation (Sec. 4.3.5). SOL stands for the pre-trained weights from SOLIDER [96] (Sec. 4.4.3).

Id	Main components of KPR						Dataset	
	PBH	MSF	Prompt		BIPO	SOL	Occluded-PoseTrack	
			$\oplus$	$\ominus$			R-1	mAP
1							84.8	74.7
2	✓						83.3 (-1.5)	72.2 (-2.5)
3	✓	✓					84.8 (+0.0)	76.1 (+1.4)
4	✓	✓	✓				88.8 (+4.0)	80.4 (+5.7)
5	✓	✓	✓	✓			90.0 (+5.2)	80.7 (+6.0)
6	✓	✓			✓		86.1 (+1.3)	76.7 (+2.0)
7		✓	✓	✓	✓		80.4 (+5.1)	89.9 (+5.7)
8	✓	✓	✓	✓	✓		**92.3** (+6.5)	**82.3** (+7.6)
9	✓	✓	✓	✓	✓	✓	90.6 (+5.8)	81.2 (+6.5)

mainly occluded samples in their test set and not in their training set. Exp. 8, in comparison with Exp. 5 and 6, demonstrates how generating these artificial multi-person occlusions at training is crucial to teach the model to rely on the input prompt to identify the intended ReID target among all candidates. The usage of the SOLIDER foundation model as pre-trained weights is illustrated in Exp. 9. Unfortunately, SOLIDER does not enhance performance on the Occ-PTrack dataset, likely because of a domain gap: SOLIDER focuses on street surveillance, while Occ-PTrack predominantly features sports images from handheld cameras. However, as demonstrated in Tab. 4.1, SOLIDER does improve performance for all other datasets.

A Prompt-Optional Method

In Figure 4.7, we analyze the test-time robustness of KPR against noisy or missing prompts on Occ-PTrack, by increasingly removing a higher percentage of keypoints from the input prompt. The red curve represents the performance of our proposed promptable KPR model, while the blue constant line illustrates the performance of the non-promptable version, which is unaffected by the proportion of keypoints removed from the input prompt. As illustrated, KPR significantly outperforms the non-promptable

Fig. 4.6 Part-attention maps for the same image but two different input prompts: KPR correctly focuses on the prompted target, whether it is the front person or the back (occluded) person. Red/pastel dots indicate negative/positive prompts. Colored heatmaps illustrates the part-attention maps of KPR (see Sec. 4.3.3), with one color per body part.

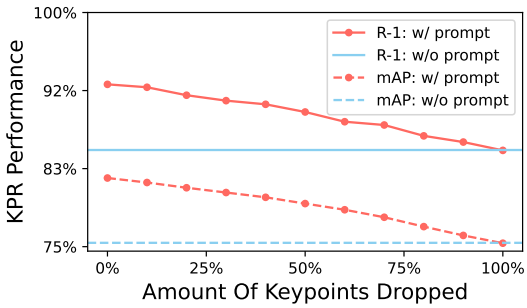


Fig. 4.7 Prompt optionality study

baseline with a full prompt and remains competitive with few or no keypoints, highlighting its prompt-optional capability.

Qualitative Assessment

Figure 4.6, Figure 4.8 and Figure 4.5 demonstrates the impact of keypoint prompting on the part-attention maps, on person retrieval, and on pose tracking. As illustrated, KPR effectively leverages the input prompts to localize the intended target, and is therefore robust to occlusions and multi-person ambiguities.

4.5 Conclusion

In this work we propose KPR, the first promptable ReID model, designed to address occluded scenarios and multi-person ambiguities arising from bounding-box-based inputs. Our model offers practical flexibility by being prompt-optional, achieving state-of-the-art performance without prompts,

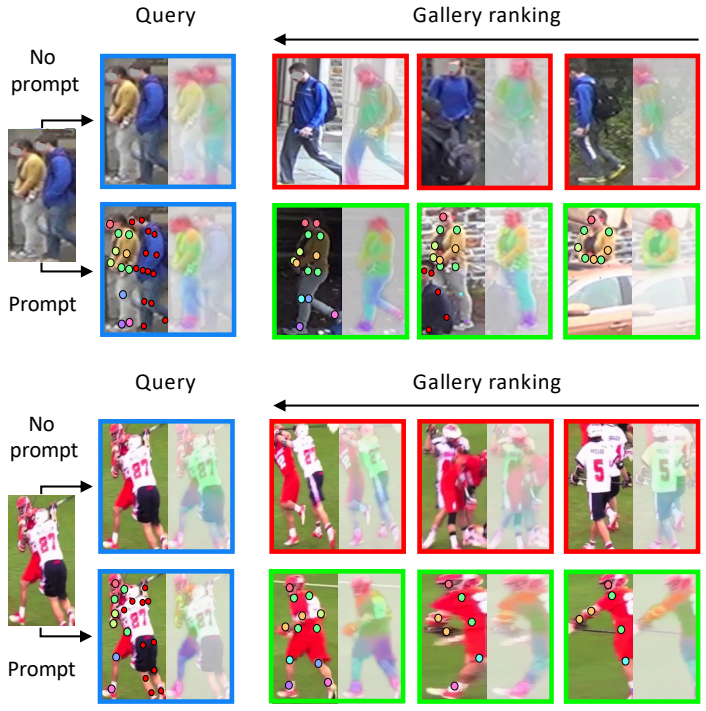


Fig. 4.8 Qualitative results of KPR on the standard person retrieval task. For a given query person, two ranking of the gallery samples are illustrated: with the input prompt enabled or disabled. The model focuses on the wrong target when no prompt is provided. Green/red borders are correct/incorrect matches. Red/pastel dots indicate negative/positive keypoint prompts. Colored heatmaps illustrates the part-attention maps of KPR described in Sec. 4.3.3, with one color per body part.

and outperforming this baseline with prompts. Furthermore, our model outperforms the state-of-the-art on three popular ReID benchmarks and on our novel Occ-PTrack dataset. We finally demonstrate KPR's potential for pose tracking in videos. Our codebase, keypoint labels and proposed dataset are released to encourage further research on this new promptable ReID paradigm.

5

CAMELTrack: Context-Aware Multi-cue ExpLoitation for Online Multi-Object Tracking

Under Review

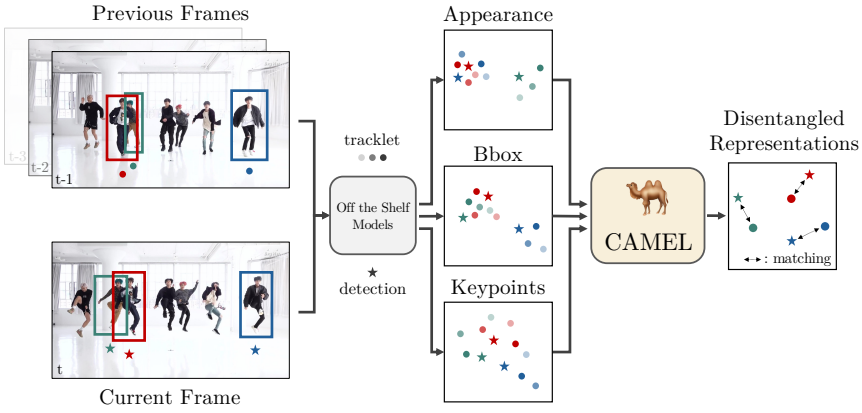


Fig. 5.1 Our proposed *CAMEL* association module for online tracking learns to produce disentangled tracklet and detection representations by combining various imperfect tracking cues.

5.1 Introduction

Multi-Object Tracking (MOT) aims to detect objects and maintain their identities across video frames, a crucial task for applications ranging from sports analytics [4, 6, 46, 102] to autonomous driving [103, 104]. In *online MOT*, decisions must be made immediately as each frame arrives, making it challenging yet crucial for real-time processing. The field is currently dominated by two paradigms: (i) SORT-based methods, and (ii) end-to-end (E2E) methods.

With the emergence of powerful object detectors [99, 105], SORT-based [40,106,107] methods, building upon the *tracking-by-detection* (TbD) paradigm, have been particularly influential. Their success stems from a *modular design*, where specialized componentsdetectors [99], re-identification models [1, 2], and motion predictors [108, 109]are independently optimized then combined through algorithmic association rules. The *association module* in SORT-based TbD pipelines, responsible for matching new detections with existing tracklets, usually encompasses three families of heuristics: (i) *tracklet representation* to aggregate frame-wise detection cues over time, (ii) *feature fusion* to combine multiple tracking cues into a single tracklet-detection cost matrix, and (iii) *multi-stage matching* to perform sequential bipartite matching operations, each utilizing distinct cues or feature fusion strategies, and operating on specific subsets of tracklets and detections. *Feature fusion*, the

most critical component of the association module, typically relies on static combinations of motion and appearance cues [40, 110–112]. However, as shown in [84, 113], cue reliability fluctuates with context particularly during occlusions, long-term associations, or when tracking visually similar targets. While some approaches attempt *context-aware feature fusion* [84, 113], their heuristic nature cannot fully capture the complex interplay between (i) the association cues and (ii) the tracked objects, suggesting the need for a more principled, data-driven approach.

To quantify the limitation of these association heuristics, we conduct an oracle-based study in Sec. 5.4.5 that reveals SORT-based methods fail to effectively leverage their strong association cues: when maintaining identical cues but replacing the association heuristic with an optimal oracle, HOTA improves by 15.5% and 8.3% on DanceTrack and SportsMOT respectively. This demonstrates substantial room for improvement within the TbD paradigm, which remains appealing due to its ability to leverage off-the-shelf models offering strong association cues. To get the best out of the TbD paradigm, we propose to learn an effective context-aware association strategies directly from data, rather than designing more sophisticated heuristics. Surprisingly, however, fully-learned association modules in online TbD remain largely unexplored. Even the transformer-based TransMOT [114], the most relevant prior work which made initial progress in this direction, still heavily relies on heuristics (as discussed further in Sec. 5.2).

To break free from these heuristics, the majority of recent literature has shifted toward the DETR-based end-to-end (E2E) paradigm, with methods like MOTR [53] offering a promising, data-driven alternative to TbD approaches.

Despite their elegant design with learned association, E2E methods face several limitations compared to SORT-based method, fully detailed in Sec. 5.2. A significant drawback is that E2E methods are designed to learn all subtasks (detection, reid, association) from scratch, forcing joint optimization of antagonistic objectives (a well-documented issue [115, 116]) while preventing the use of specialized external models. These fundamental limitations consequently require substantial training data and computational resources, typically several days of training on 8 GPUs.

Given the limitations of both E2E and SORT-based methods, we bridge the gap between the two paradigms by proposing **CAMEL**, a novel association module for **C**ontext-**A**ware **M**ulti-**C**ue **E**xp**L**oitation that replaces traditional SORT-like association heuristics with a unified trainable architecture. CAMEL's compact and minimalist architecture consists of: (i) a set of

Paradigm	Association	Methods	HF	OM	LC
TbD	Heuristic	SORT-based [40, 107]	✗	✓	✓
	Hybrid	*TransMOT* [114]	✗	✓	✓
	Learned	*CAMEL* (ours)	✓	✓	✓
DbT	E2E MOT	*MOTR* [53]	✓	✗	✗

Table 5.1 Comparison of popular association paradigms and methods for Online MOT. HF stands for Heuristic-Free association, OM refers to the ability to use Off-the-shelf Models, and LC denotes Low training Compute. We refer readers to Figure 2.15 in the Background section for a more comprehensive taxonomy of existing tracking paradigms.

Temporal Encoders (TE) that aggregate each tracking cue into tracklet-level representations, and (ii) a Group-Aware Feature-Fusion Encoder (GAFFE) that jointly transforms all cues into unified disentangled representations for each tracklet and detection. As illustrated in Fig. 5.1, CAMEL properly discriminates matching tracklets and detections despite occlusions or similar-looking targets, by dynamically balancing multiple imperfect association cues. This capability stems from its context-aware processing, that accounts for interactions between targets and the relative discriminativeness of each cue. Our resulting *heuristic-free online TbD* tracker, **CAMELTrack**, achieves state-of-the-art performance on five popular MOT benchmarks.

Overall, we summarize our contributions as follows:

- We propose CAMEL which, to our knowledge, represents the first fully-learned and cue-agnostic association module for TbD pipelines, designed without bells and whistles. CAMELTrack runs at 13 FPS, which is faster than previous transformer-based trackers.
- We introduce an efficient Association-Centric Training, requiring under an hour on a single GPU, whereas E2E methods typically need days on multiple GPUs.
- We show that learned association with off-the-shelf models outperforms both E2E and SORT-based methods across five challenging benchmarks, effectively combining the strengths of both paradigms.

We release our framework and models weights to encourage further research on learned TbD association modules.

5.2 Related Work

In this section, we review key MOT approaches related to our work, focusing on online methods.

Heuristic-based Tracking-by-Detection. The dominant paradigm in MOT has been tracking-by-detection (TbD), with many methods building upon SORT [106]. These approaches focus on developing sophisticated association heuristics [40, 107, 111, 112], or stronger motion modeling [108, 109, 117–122] and re-identification [84, 123–125]. Distinct SORT-based methods primarily differ in their hand-crafted rules for association across three key components: (i) *Tracklet Representation*: common approaches include mean [52] or exponential moving averages of detection features [100, 101], or minimal distance to a feature bank [40]. GHOST [84] provides a comprehensive analysis of various "proxies" for computing the distance between a tracklet and a single detection, including the "Exponential Moving Average Feature Vector", "Median Feature Vector", "Last Frame Feature Vector", among others. (ii) *Feature Fusion*: methods range from weighted averaging of motion and appearance cues [84] or additional cues [110, 125], to adaptive weighting schemes [113] and threshold-based gating [111, 112]. GHOST [84] also conducts an extensive study examining how different "Motion Weight" values (weighting factors combining motion and appearance cost matrices) impact tracking performance across various datasets. (iii) *Multi-stage Matching*: trackers employ either single-stage [111] or cascaded matching [40], filtering candidates based on confidence scores [107] or track age [40], while using different cue at each stage. As described in Sec. 5.1, Multi-stage matching involves computing distinct association cost matrices at each stage, using carefully selected subsets of active tracklets and detections (filtered by detection confidence or tracklet age). Each stage employs the Hungarian algorithm for bipartite matching, with unmatched tracklets/detections being processed in subsequent stages.

The majority of recent state-of-the-art methods typically implement a two-stage approach: an initial matching stage using custom heuristics (often incorporating ReID features), followed by a motion-based stage using IoU between Kalman Filter predicted bounding boxes and current detections, following SORT's [106] original design. For example, DeepSORT performs multiple cascade matching stages using ReID features, processing tracklets in order of age, before concluding with SORT's standard Kalman Filter association stage.

Our method take a different direction and replaces these heuristics for

data association with a unified trainable architecture, that better leverages all available tracking cues to produce context-aware disentangled representations to be matched in a single stage.

Tracking-by-Detection with Learned Association. While some previous works have explored data-driven tracking through graph networks [126] or transformers [127, 128], most operate offline, with only a few pioneering works attempting to integrate learned components into online TbD pipelines [114, 129–131]. Our proposed CAMELTrack falls within this category of MOT methods, with TransMOT [114] being the closest prior work. TransMOT [114] introduces a spatial-temporal encoder for tracklet representation and a transformer for feature fusion, but relies on a hand-crafted multi-stage matching pipeline where the learned components are only used in the second stage, while the first and third stages remain purely based on IoU and re-identification (ReID) heuristics. Moreover, CAMEL has a fundamentally different architecture compared to TransMOT, that uses a graph-inspired approach with $N \times M$ edge tokens, while CAMEL follows a metric learning approach, encoding $N + M$ tracklets and detections into a shared disentangled latent space. While these works represent initial steps toward learned association, they still remain dependent on heuristics. In contrast, our approach makes a decisive break from hand-crafted rules by introducing a completely trainable association module

Online Detection-by-Tracking. Recently, end-to-end (E2E) methods [53, 104, 115, 116, 132–137] following the Detection-by-Tracking (DbT) paradigm [52] have emerged as a promising, heuristic-free alternative to TbD approaches. Building upon DETR [105] architecture, these methods jointly learn object detection and association, using track queries to re-detect past objects across frames. Despite their elegant design that learns association in a data-driven way similar to our approach, E2E still struggle to reach SoTA performance on a wide range of datasets. This is because E2E methods face several limitations: (i) their detector-centric multi-frame training with short time windows struggles with long-term associations [138], (ii) they lack TbD's modular ability to leverage specialized external models (e.g., ReID, motion, ...) [116], (iii) the inherent conflict between detection and association objectives [115] in a shared model limits their overall performance and (iv) they require extensive training data and computational resources to achieve competitive performance (typically a few days on 8 GPUs [53]). Our approach is fundamentally different from E2E methods and does not fall into that category. Indeed, it focuses solely on learning an association strat-

egy, requiring an order of magnitude less training compute, and maintains TbD's ability to leverage off-the-shelf detection, motion, and ReID models.

5.3 Methodology

In this section, we detail CAMELTrack, our proposed online tracking method. We first provide an overview of the complete tracking pipeline in Sec. 5.3.1. In Sec. 5.3.2, we then detail CAMEL, our trainable Context-aware Multi-cue ExpLoitation module that learns tracklet-detection association directly from data. Finally, we describe our association-centric training scheme designed to create challenging association scenarios in Sec. 5.3.3.

5.3.1 CAMELTrack Pipeline

Our multi-object tracking pipeline, CAMELTrack, follows the standard online tracking-by-detection paradigm, processing each incoming frame through four sequential steps: (i) object detection, (ii) cue extraction, (iii) tracklet-detection association via our CAMEL module, and (iv) tracklet life cycle management. The following paragraphs detail one complete iteration of this process, that is illustrated in Fig. 5.2.

Detection. We first process the incoming video frame with timestamp t^{cur} with an object detector to obtain a set of detections $\mathcal{D}$, where each detection $d^{t^{\text{cur}}}$ is represented by a bounding box and its confidence score.

Cue Extraction. For each detection in $\mathcal{D}$, we extract multiple complementary cues to guide the association process, as a single cue is often insufficient for reliable tracking. The bounding box coordinates and confidence score constitute the first cue c_0, while K additional cues $\{c_k\}_{k=1}^{K}$ are extracted by specialized off-the-shelf models. In this work, we employ *re-identification features* and *pose keypoints* as additional cues to complement the object location c_0. However, our CAMEL association module can ingest any type and number of input cues, enabling easy integration of additional domain-specific information (e.g., plate numbers for vehicle tracking). Each detection d is thus characterized by its complete cue set, i.e. $d^{t^{\text{cur}}} = \{c_k^{t^{\text{cur}}}\}_{k=0}^{K}$.

Association with CAMEL. The association step objective is to match M existing tracklets $\mathcal{T}$ with N active detections $\mathcal{D}$ from the current frame. We refer to all tracklets and detections considered for association as the set of *active objects* $\mathcal{A} = \mathcal{T} \cup \mathcal{D}$. Each tracklet in $\mathcal{T}$ represents a unique tracked object and is composed of a sequence of detections $d^{t^{\text{start}}:t^{\text{end}}} = [d^{t^{\text{start}}}, \ldots, d^{t^{\text{end}}}]$,

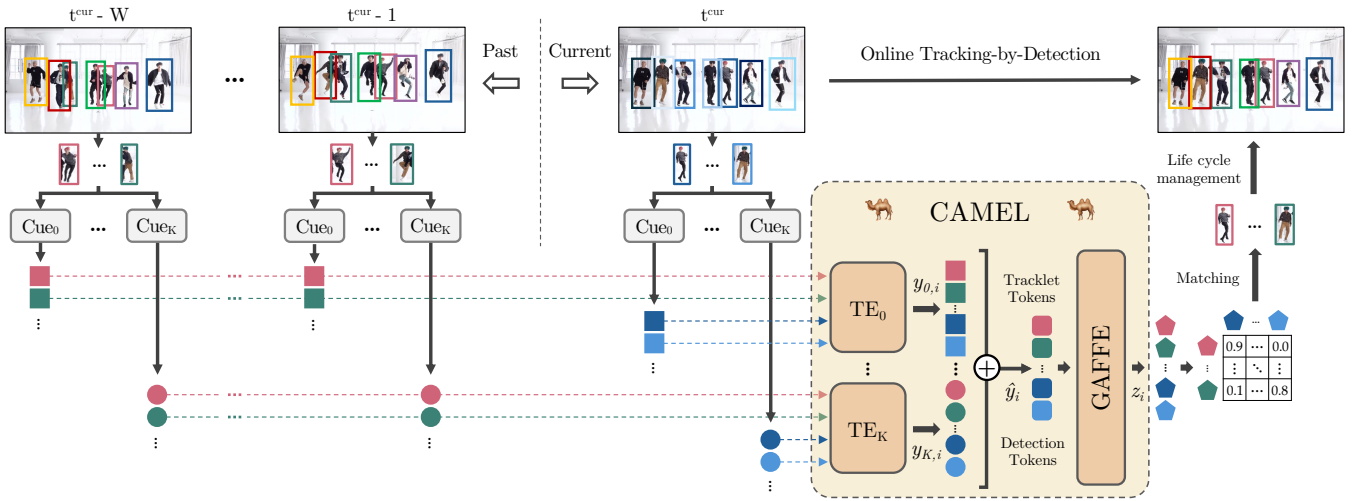


Fig. 5.2 Architecture overview of CAMELTrack, our online tracking-by-detection pipeline that operates in three steps: (i) object detection, (ii) cue extraction, (iii) single-stage association using our trainable CAMEL module, and (iv) tracklet life cycle management. CAMEL processes the various imperfect cues through two stages: First, the Temporal Encoder (TE) aggregates each cue into tracklet-level representations. Second, the Group-Aware Feature Fusion Encoder (GAFFE) embeds all detection and tracklet cues into a unified discriminative embedding space. The resulting disentangled tracklets and detections representations are finally paired through bipartite matching.

where t^{start} and t^{end} indicate respectively the frame indices of the first and last detection in the tracklet. For each active tracklet, we maintain a feature bank storing the cues of its W most recent detections allowing CAMEL to leverage a rich history of cues to counter potential noise in individual detections or id switches resulting from association errors. CAMEL is the core contribution of our work. It takes as input all active tracklets $\{d_i^{t_i^{\text{start}}:t_i^{\text{end}}}\}$ for $i \in \mathcal{T}$ and detections $\{d_j^{t^{\text{cur}}}\}$ for $j \in \mathcal{D}$, and outputs a single discriminative embedding z per active object (detection and tracklet) in a shared latent space, where matched/unmatched pairs are localized close/far to each other. The context-aware architecture of CAMEL is detailed in Sec. 5.3.2, and its training procedure in Sec. 5.3.3. Finally, CAMEL's disentangled representations are used to compute a cost matrix $C \in \mathbb{R}^{M \times N}$, where each entry $c_{i,j} = ||z_i - z_j||_2$ measures the Euclidean distance between the normalized embeddings z_i of tracklet i and z_j of detection j. The final assignment is then obtained through bipartite matching with the Hungarian algorithm. Any pair whose cost exceeds a specified threshold is left unmatched.

Life Cycle Management. CAMELTrack manages tracklet life cycles through a standard scheme: first, low-confidence detections are filtered out before association. Next, each matched detection extends its assigned tracklet by adding new cues to its feature bank. Unmatched high-confidence detections initialize new tracklets, while unmatched tracklets are temporarily paused and eventually terminated if they remain unmatched for an extended period.

5.3.2 Our CAMEL Architecture

In this section, we detail CAMEL, our trainable association module for Context-Aware Multi-Cue ExpLoitation. As introduced in Sec. 5.3.1, CAMEL takes as input all cues from all active objects $\mathcal{A} = \mathcal{T} \cup \mathcal{D}$, where $\mathcal{T}$ includes the M existing tracklets and $\mathcal{D}$ the N current detections, and outputs their unified representations in a disentangled space. As a result, objects with same/distinct identities are embedded close/far to each other. As introduced in Sec. 5.1, CAMEL replaces the three key heuristics traditionally used in SORT-based association modules *tracklet representation*, *feature fusion*, and *multi-stage matching* with a unified trainable architecture built upon two transformer-based components: the Temporal Encoder (TE) and the Group-Aware Feature Fusion Encoder (GAFFE). First, the Temporal Encoder (TE) performs intra-object self-attention to aggregate detection-level cues into robust tracklet-level representations. Next, the Group-Aware

Feature Fusion Encoder (GAFFE) fuses multiple imperfect but complementary cues into a unified representation for each object. Through inter-object self-attention, it maximizes discriminativeness between objects of different identities while enhancing similarity among objects of the same identity. Both TE and GAFFE are introduced separately in the next subsections. Finally, the need for *multi-stage matching* naturally disappears as CAMEL processes all tracklets, detections, and cues at once, to perform association in a single unified stage.

Temporal Encoder (TE). Each active object is processed by K Temporal Encoders, each TE_k tackling a specific cue type and having a dedicated set of weights. For a given active object $i \in \mathcal{A}$ and cue k, the Temporal Encoder TE_k processes the temporal sequence $c_{k,i}^{t^{\text{start}}:t^{\text{end}}} = [c_{k,i}^{t^{\text{start}}}, \ldots, c_{k,i}^{t^{\text{end}}}]$ as follows. First, each cue $c_{k,i}^{t}$ in this sequence undergoes a linear transformation to produce a token $x_{k,i}^{t}$. This critical step embeds low dimensional cues like bounding boxes into a high dimensional feature space. Next, each token $x_{k,i}^{t}$ is augmented with a sinusoidal positional encoding (PE) that encodes its relative temporal position, i.e., age, compared to the current frame timestamp t^{cur},

$$\tilde{x}_{k,i}^{t} = x_{k,i}^{t} + \text{PE}(t^{\text{cur}} - t). \tag{5.1}$$

Then, a learned `[CLS]` token is prepended to the sequence of tokens $\tilde{x}_{k,i}^{t_i^{\text{start}}:t_i^{\text{end}}}$, and the resulting sequence is processed by a shallow multi-layer transformer encoder [139].

Finally, the encoded CLS token serves as the output of the TE, providing a single temporal representation $y_{k,i}$ of cue k for object i,

$$y_{k,i} \leftarrow \text{TE}_k([\texttt{[CLS]}, \tilde{x}_{k,i}^{t_i^{\text{start}}}, \ldots, \tilde{x}_{k,i}^{t_i^{\text{end}}}]). \tag{5.2}$$

Both tracklets in $\mathcal{T}$ and detections in $\mathcal{D}$ undergo temporal encodingeven if detections are sequences length of oneto ensure all cues are embedded in the same latent space for further processing by GAFFE.

Group-Aware Feature Fusion Encoder (GAFFE). This module receives as input the temporally encoded tokens $y_{k,i}$ produced by the Temporal Encoders, where each token corresponds to a different cue of each active object $i \in \mathcal{A}$. GAFFE processes these tokens in two stages to produce a single discriminative embedding per object.

In the first stage, each cue-specific token $y_{k,i}$ is linearly projected into a higher-dimensional space. The projected tokens are then fused through

summation to generate a single multi-modal token $\hat{y}_i$ per active object,

$$\hat{y}_i = \sum_{k=0}^{K} \text{Linear}_k(y_{k,i}). \tag{5.3}$$

In the second stage, the resulting sequence of $N + M$ multi-modal tokens $\hat{y}_i$ is processed through a shallow multi-layer transformer encoder [139], that performs group-aware inter-object self-attention,

$$\{z_i\} \leftarrow \text{GAFFE}(\{\hat{y}_i\}), \quad \forall i \in \mathcal{A}. \tag{5.4}$$

These resulting embeddings $\{z_i\}$ are the final, disentangled representations of each active object, which are then used for matching as detailed in Sec. 5.3.1.

5.3.3 Association-Centric Training

Existing end-to-end (E2E) methods employ a *recursive multi-frame training scheme* [53,116], where the model processes a short video sequence frame-by-frame to learn detection and association jointly. In contrast, our proposed Association-Centric Training (ACT) strategy decouples association from both detection and cue extraction, and works as follows. First, we generate an image-free training set by (i) running an off-the-shelf detector on all training sequences, (ii) assigning each detection the label of its IoU-closest ground-truth, (iii) extracting all required cues (e.g., re-identification, pose). Then, during training, we sample from our pre-generated set to build batches of B training samples. Each training sample corresponds to one input of CAMEL and models a single association scenario with P tracklet-detection pairs. A single scenario is constructed by choosing a random frame, collecting all detections from that frame along with tracklets from previous frames. We repeat this process with frames from distinct videos until P pairs are sampled. This cross-video sampling to generate artificial association examples increases training diversity and empirically yields more stable training and faster convergence. We further enrich training by applying three data augmentations to generate more challenging and diverse association scenarios: (i) *detection identity swapping*, (ii) *detection dropout*, and (iii) *cue dropout*. Finally, we employ the InfoNCE loss [140] as a training objective to minimize/maximize distances between detection-tracklet pairs of same/different identities.

ACT offers two key advantages over recursive training strategies. First,

E2E methods are computationally constrained to short sequences due to their heavy image processing architectures. In contrast, our lightweight processing of pre-computed features enables efficient modeling over large time windows, thereby improving long-term tracking. Second, ACT's data augmentations generate synthetic training samples that model diverse challenging scenarios: occlusions, similar-looking targets, scene re-entries, noisy features, and detection errors. As demonstrated in Sec. 5.4.5, exposure to these hard examples significantly improves performance.

5.4 Experiments

5.4.1 Datasets and Metrics

We evaluate CAMELTrack on three datasets. **DanceTrack** [45] features similar dancers in complex choreographies, while **SportsMOT** [46] focuses on players in team sports. These benchmarks, extensively studied by recent methods, present complementary tracking challenges while comprising decent training, validation and testing sets. Additionally, **PoseTrack21** [48] serves as a diverse real-world testbed where we demonstrate our method's modularity through effective keypoint integration. We exclude **MOT17** [44] from our main evaluation despite its popularity, because we believe this dataset is not suited for data-driven MOT approaches; we discuss this further at the end of the chapter.

For evaluation, we primarily rely on HOTA [51], that combines two complementary metrics, Detection Accuracy (DetA) and Association Accuracy (AssA). We also report the standard MOTA [49] and the IDF1 [50], the latter focusing on identity preservation across long trajectories. We focus our analysis on association-related metrics (AssA & IDF1) as they directly evaluate our contribution's impact, independently of detection quality.

5.4.2 Implementation details

We use the YOLOX [99] detector provided by DiffMOT [108] for DanceTrack and SportsMOT. For PoseTrack21, we train our own YOLOX. For tracking cues, we leverage dataset-specific BPBReid [1] models for appearance, off-the-shelf RTMPose [141] for DanceTrack/SportsMOT pose estimation, and use HRFormer [142] for PoseTrack21. Our tracking pipeline is implemented with TrackLab [10]. Our model employs 4-layer, 8-head transformer encoders for both TEs and GAFFE, for a total 42.6M parameters. Training occurs over 10 epochs using AdamW [143] with a 1e$-$4 learning rate

and a batch size of $B = 8$. A training sample has $P = 32$ detection-tracklet pairs. We first pretrain the TEs independently before jointly optimizing with GAFFE. Training takes one hour on a single consumer-grade GPU, while the inference association step runs in 18ms. For DanceTrack/SportsMOT/PoseTrack, detection confidence thresholds are set to 0.4/0.1/0.3, minimal confidence for tracklet initialization at 0.9/0.4/0.4 and tracklet-detection similarity threshold at 0.1/0.1/0.45. We employ a feature bank of $W = 50$ frames and maintain paused tracklets for 150 frames.

5.4.3 Inference Speed

Without specific optimization, our pipeline operates at 14 FPS on DanceTrack. Individual components achieve: YOLOX-X (38 FPS), RTMPose (91 FPS), BPBReid (63 FPS), and CAMELTrack (54 FPS). CAMELTrack is therefore faster than other transformer-based solutions like MOTR [53] (7 FPS), MOTRv2 [115] (7 FPS), and TransMOT [114] (9 FPS), while being slower than lightweight TbD methods like DiffMOT [108] (23 FPS).

5.4.4 Comparison to State-of-the-Art

Our method establishes new state-of-the-art performance across all evaluated benchmarks, surpassing both end-to-end (E2E) methods [53, 116, 134] that traditionally dominate DanceTrack, and SORT-based approaches [84, 108, 110, 113] that excel on SportsMOT. Furthermore, CAMELTrack outperforms existing methods on PoseTrack21 [48, 52, 100, 101, 144], demonstrating its natural extension to pose tracking.

DanceTrack. As depicted in Tab. 5.2, E2E methods [53, 116, 134] dominate this benchmark, outperforming existing SORT-based methods [84, 107–110, 113].

The poor performance of these SORT-based methods can be attributed to DanceTrack's challenging scenarios — similar-looking dancers executing complex movements with frequent occlusions — which yield unreliable motion and appearance cues, as demonstrated by our oracle-based study in Sec. 5.4.5. Heuristic-based association is inherently more sensitive to such unreliable inputs: incorrect associations therefore occur, progressively degrading tracklets' representations and cascading into even more tracking errors. While Hybrid-SORT [110], attempts to address these issues by introducing three additional cues, it still remains limited by a static feature fusion. In contrast, our data-driven association bridges the performance gap with E2E methods by learning to leverage each cue's discriminative power.

Method	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑
End-to-End MOT					
MOTR [53]	54.2	40.2	73.5	51.5	79.7
MOTIP [116]	67.5	57.6	79.4	72.2	90.3
MeMOTR [134]	68.5	58.4	80.5	71.2	89.9
Heuristic Association					
ByteTrack [107]	47.7	32.1	71.0	53.9	91.3
OC-SORT [109]	55.1	38.0	80.3	54.2	89.4
GHOST [84]	56.7	39.8	81.1	57.7	89.6
Deep OC-SORT [113]	61.3	45.2	82.2	61.5	92.3
DiffMOT [108]	62.3	48.8	**82.5**	64.0	**92.7**
Hybrid-SORT [110]	65.7	-	-	67.4	91.8
Learned Association					
CAMELTrack	66.1	54.0	81.1	71.1	91.4
w/ keypoints	**69.3**	**58.9**	81.8	**74.9**	91.4

Table 5.2 Comparison on DanceTrack [45] test set. For fair comparison, we only report methods trained exclusively on DanceTrack.

Similar to our approach, MeMOTR [134] and MOTIP's [116] success can be attributed to their learned association and robust tracklet representation.

Finally, previous attempts [45] at leveraging keypoints achieved only marginal gains (+0.4% HOTA), likely due to hand-crafted rules' limitations in exploiting this rich information. In contrast, our method yields significant improvements (+3.2% HOTA), surpassing E2E performance.

SportsMOT. As reported in Tab. 5.3, SORT-based [46, 108, 117, 145] methods dominate SportsMOT's leaderboard, outperforming E2E solutions [116, 134]. This success can be attributed to appearance and motion cues being more reliable on SportsMOT than on DanceTrack. For instance, even though players wear similar team uniforms, our ablation study in Sec. 5.4.5 demonstrates that appearance remains a very effective cue for sports tracking. The effectiveness of appearance cues particularly benefits TbD methods, as their dedicated ReID models capture object appearance better than E2E track-queries.

On the other hand, we outperform SORT-based methods for similar reasons to DanceTrack. Our Association-Centric Training exposes the model to long-term associations, which improves handling of scene re-entries. Overall, CAMELTrack achieves significant improvements (+3.2% HOTA) over prior state-of-the-art methods. However, unlike DanceTrack, keypoints degrade performance on SportsMOT, likely due to more distant viewpoints

Method	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑
End-to-End MOT					
MeMOTR [134]	70.0	59.1	83.1	71.4	91.5
MOTIP	71.9	62.0	83.4	75.0	92.9
Heuristic Association					
ByteTrack [107]	62.1	50.5	76.5	69.1	93.4
OC-SORT [109]	68.1	54.8	84.8	68.0	93.4
MotionTrack [117]	74.0	61.7	88.8	74.0	96.6
MixSORT [46]	74.1	62.0	88.5	74.4	96.5
DiffMOT [108]	76.2	65.1	**89.3**	76.1	**97.1**
Deep-EIoU [145]	77.2	67.7	88.2	79.8	96.3
Learned Association					
CAMELTrack	**80.4**	**72.8**	88.8	**84.8**	96.3
w/ keypoints	80.3	72.6	89.0	**84.8**	96.4

Table 5.3 Comparison on SportsMOT [46] test set.

resulting in noisy pose estimation.

PoseTrack21. As reported in Tab. 5.4, current methods can be categorized into two settings: methods with private detections extending traditional MOT frameworks [52, 100, 101], and those using public detections with a custom pose-aware tracking [48, 144]. Different from SportsMOT and DanceTrack, PoseTrack covers diverse real-world scenarios with dramatic camera motion, viewpoint changes, and motion blur, making detection and association challenging.

Using public detections from [48, 144], CAMEL establishes new state-of-the-art performance with significant gains (+3.8% AssA) through an effective fusion of appearance, motion and pose cues. With stronger private detections, CAMEL achieves even larger gains (+7.7% HOTA).

5.4.5 Ablation Studies

We conduct extensive experiments in Tab. 5.5 on SportsMOT and DanceTrack validation sets to analyze CAMEL's design. Our study evaluates three key aspects: (i) Temporal Encoders versus standard tracklet representations heuristics (Exp. 1-5), (ii) our Group-Aware Feature Fusion Encoder (Exp. 6-8), and (iii) our complete architecture (Exp. 9-10). Additionally, we design oracle experiments (Exp. 11-12) to establish performance upper bounds.

Temporal Encoders vs. Heuristics. These experiments compare our TE with standard heuristics using different cues. Regarding Re-ID features,

Method	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑
Public Detection Setting					
CorrTrack [48]	57.0	64.2	51.3	66.5	52.0
GAT [144]	58.4	66.9	**51.8**	-	**55.3**
CAMELTrack	**58.7**	**70.7**	50.0	**67.8**	51.7
Private Detection Setting					
TRMOT [100]	46.9	55.0	40.9	57.3	47.2
FairMOT [101]	53.5	61.5	47.4	63.2	56.3
Tracktor++ [52]	58.3	65.4	52.7	69.3	59.5
CAMELTrack	**66.0**	**73.8**	**59.9**	**76.0**	**67.5**

Table 5.4 Comparison on PoseTrack21 [48] validation set.

TE consistently outperforms the Exponential Moving Average (EMA) (Exp. 1-2). This improvement is particularly noteworthy as appearance is a weak cue on DanceTrack but highly discriminative on SportsMOT. Similarly for bounding box cues, TE outperforms Kalman Filter's (KF) predictions on DanceTrack's erratic movements and frequent occlusions (Exp. 3-4). On the other hand, KF effectively captures the more predictable player trajectories in SportsMOT. Pose keypoints provide complementary information, especially for distinguishing dancers during occlusions, but show no improvement over bounding box tracking on SportsMOT, likely due to noisy estimates from distant views (Exp. 5).

Feature Fusion Analysis. We evaluate GAFFE's learned dynamic feature fusion against static rules. The baseline using equal weights for motion and appearance features (Exp. 6) shows no significant gain over using cues independently, and sometimes even decreases performance. Adding GAFFE for group-aware feature fusion (Exp. 7) yields consistent improvements, demonstrating the benefits of a learned approach. Using both temporal and group-aware encoding (Exp. 8) provides additional gain, with DanceTrack particularly benefiting from this combination.

Complete Architecture and Training. Ablating data augmentation (Exp. 9) during our association-centric training significantly degrades performance, demonstrating the importance of training on diverse scenarios. Our final architecture with pose information (Exp. 10) achieves the strongest results on DanceTrack while showing no improvements on SportsMOT due to its distant camera setup.

These results demonstrate CAMEL's key strengths in two key aspects:

Exp	Features			GAFFE	DA	DanceTrack		SportsMOT	
	App	Bb	Kp			HOTA↑	IDF1↑	HOTA↑	IDF1↑
1	EMA	-	-	✗	✗	49.9	48.0	76.0	80.8
2	TE	-	-	✗	✓	52.4 (+2.5)	52.6 (+4.6)	79.2 (+3.2)	84.7 (+3.9)
3	-	KF	-	✗	✗	54.3	56.3	72.1	72.6
4	-	TE	-	✗	✓	54.7 (+0.4)	57.4 (+1.1)	71.1 (-1.0)	75.4 (+2.8)
5	-	-	TE	✗	✓	56.0	59.5	71.3	75.7
6	EMA	KF	-	✗	✗	54.3	57.2	75.8	80.7
7	EMA	KF	-	✓	✓	56.5 (+2.2)	58.2 (+1.0)	79.4 (+3.6)	85.1 (+4.4)
8	TE	TE	-	✓	✓	62.4	65.7	81.8	87.9
9	TE	TE	TE	✓	✗	61.0	64.9	78.2	83.7
10	TE	TE	TE	✓	✓	65.1 (+4.1)	70.5 (+5.6)	81.9 (+3.7)	88.5 (+4.8)
11	Oracle Feature Fusion (KF & EMA)					69.8	74.7	84.1	91.0
12	Oracle Association					86.1	98.2	90.8	99.4

Table 5.5 Ablation study on the validation set of each dataset. App stands for appearance embeddings, EMA for exponential moving average, Bb for bounding box, KF for Kalman Filter's predicted box, Kp for keypoints, and DA for the Data Augmentation.

(i) better temporal encoding through TE's robust representations, and (ii) improved discrimination via GAFFE's dynamic weighting, enabling what we refer to as *group-aware feature fusion*.

Analyzing TbD Association through Oracles. Two oracle experiments have been designed to study the limitations of Tracking-by-Detections (TbD) heuristic-based association and evaluate the discriminative power of motion and appearance cues. First, we design a Feature Fusion Oracle (Exp. 11) that linearly combines motion and appearance cues so as to result in a cost matrix that maximizes the association accuracy. This oracle reveals two key insights: (i) motion and appearance are two strong and highly complementary cues for tracking, but (ii) the significant gap with standard fusion methods (Exp. 6) reveals that static heuristics fail to fully leverage their discriminative power. Second, the Association Oracle (Exp. 12), which matches each detection to its IoU-closest ground truth track, establishes an absolute upper bound on association performance with detection quality as the only limitation. The performance gap between Feature Fusion and Association Oracles varies significantly across datasets: the small gap on SportsMOT indicates reliable tracking cues, while the large gap on DanceTrack reveals the need for stronger cues in such challenging scenarios. Overall, we find

encouraging the results showing that *our learned association strategy contributes to bridge the gap towards oracle performance* (Exp. 10-11 achieve close performances).

5.4.6 Qualitative Analysis of Latent Representations

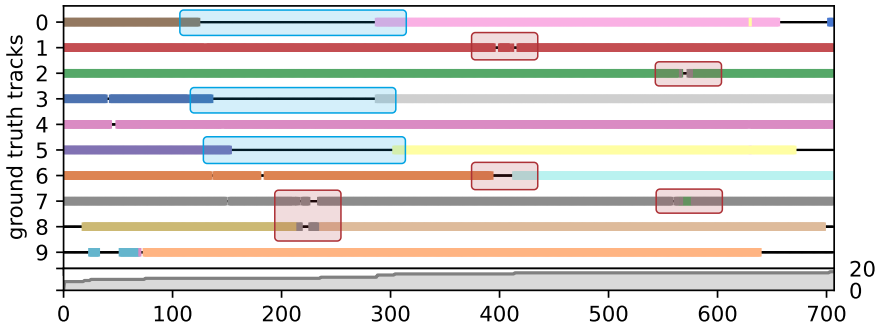
To illustrate CAMEL's cue disentanglement capabilities, we analyze similarity distributions between tracklet-detection pairs and latent space structure using t-SNE [146]. We compare CAMEL's output embeddings with standard heuristic cues: Kalman Filter (KF) for motion and Exponential Moving Average (EMA) of Re-ID embeddings for appearance.

Analysis of Similarity Distributions. Fig. 5.5 compares the similarity distributions between tracklet-detection pairs that share the same identity (positive) versus pairs with different identities (negative), for standard motion/appearance cues and CAMEL's outputs. While KF motion cues effectively discriminates most positive pairs from negative ones, a significant portion still exhibits incorrect low IoU values. This limitation is particularly evident on DanceTrack, where negative pairs frequently overlap with positive ones, highlighting KF's weakness. Moreover, appearance alone lacks discriminativeness, as evidenced by the non-negligible overlap between positive and negative pairs, especially on DanceTrack. In contrast, CAMEL's output embeddings effectively discriminate positive from negative pairs, demonstrating a successful cue disentanglement.

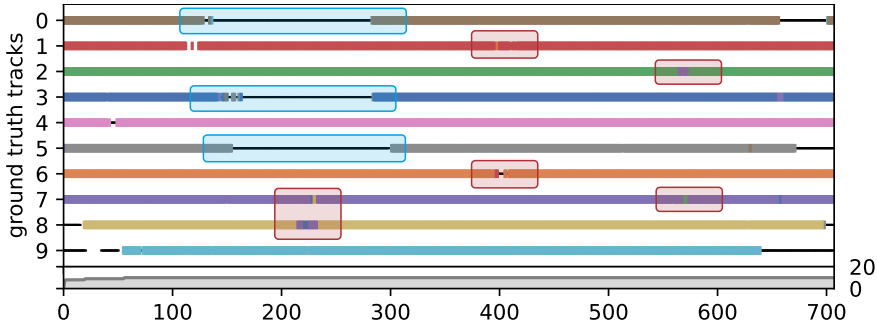
Latent Space analysis through t-SNE. Fig. 5.4 illustrates the t-SNE representations of motion, appearance, and CAMEL outputs on a short sequence with heavy occlusions, where tracklets (darker) and detections (lighter) are colored by identity. Motion embeddings organize into identity clusters but show significant overlap during occlusions, while appearance features achieve better but incomplete separation. On the other hand, CAMEL outputs form distinct identity clusters with minimal overlap, demonstrating an effective combination and disentanglement of these complementary cues.

5.4.7 Qualitative Results

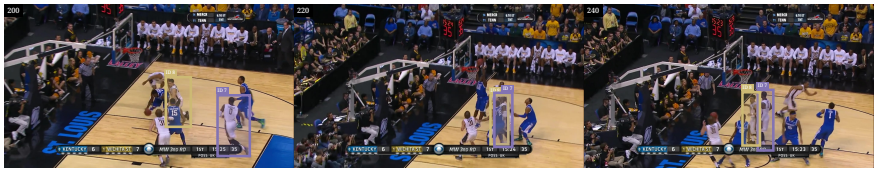
Fig. 5.3 compares CAMELTrack with the competitive DiffMOT [108] using the same detections on a challenging SportsMOT sequence featuring scene re-entries and heavy occlusions. This figure illustrates their tracking performance through a timeline where ground truth tracks are represented with horizontal lines, and identities with different colors. For both methods, a



(a) DiffMOT [108]

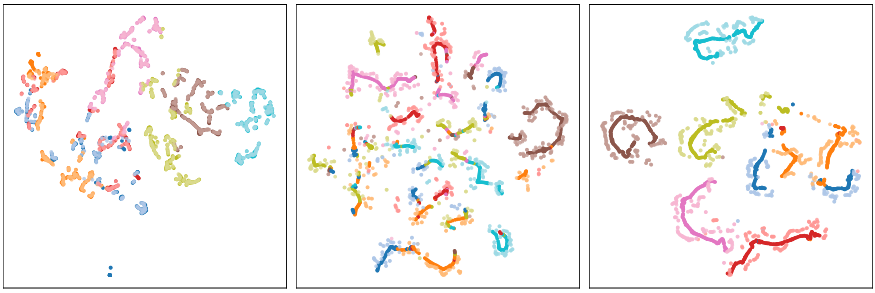


(b) CAMEL



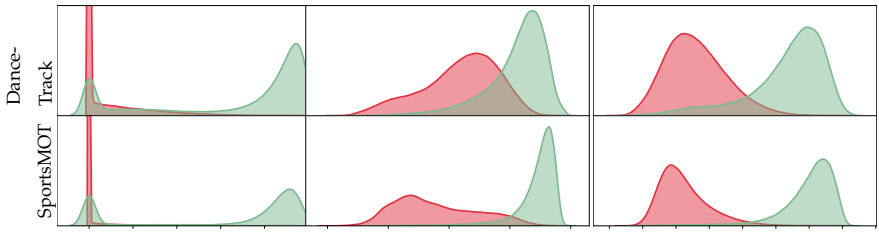
(c) Occ. Example

Fig. 5.3 Visualization of tracking results on *v_00HRwkvvjtQ_c007* from SportsMOT. Ground truth tracks are depicted with horizontal lines, while colors indicates predicted identities. Blue zones highlight scene re-entries and red zones show occlusions. Frames where the ground truth identity has left the scene are represented with a black line, and missing predictions are left blank. Bottom gray plots show cumulative tracked identities over time. (c) Frames from the highlighted occlusion between ids 7 & 8 around frame 200.



(a) Motion **(b)** Appearance **(c)** CAMEL output

Fig. 5.4 t-SNE on the first 150 frames of *dancetrack0019*. Each identity is assigned a unique color, with light/dark shades indicating detections/tracklets respectively.



(a) Motion **(b)** Appearance **(c)** CAMEL output

Fig. 5.5 Similarity distributions between positive and negative tracklet-detection pairs. (a) IoU between KF's predictions and detections. Cosine distances between : (b) EMA tracklet and detection ReID embeddings (c) pairs of CAMEL's output embeddings.

cumulative plot shows the total number of unique identities created over time.

Both methods show distinct behaviors during scene re-entries: while DiffMOT generates new identities, CAMEL successfully recovers known ones through its feature bank, as shown by the lower slope in the cumulative identity plot. Similarly, during occlusions, both methods initially make identity switches, but CAMEL recovers from these errors while DiffMOT propagates them forward.

5.5 Conclusion

We introduced CAMEL, a novel learned association module that replaces traditional hand-crafted rulestracklet representation, feature fusion, and multi-stage matchingwith a unified trainable architecture. With our state-of-the-art performance, we hope to reaffirm TbD as a strong paradigm for online tracking and encourage a shift from heuristics-based association toward learned approaches. We view this work as a first step to demonstrate the potential of fully learned TbD association modules and release our codebase to foster future research in this direction. Building upon CAMEL, future work could explore more sophisticated data augmentation, training objectives, neural architectures, or extend the learning paradigm to additional components like tracklet life cycle management.

5.6 Additional Discussion about MOT17

MOT17 remains a well-established dataset within the MOTChallenge benchmark and has historically served as a primary evaluation platform for multi-object tracking. However, as highlighted by recent works [53, 114–116, 134], several fundamental limitations make MOT17 particularly unsuitable for evaluating learned association approaches like CAMEL. Hereafter, we discuss important limitations of MOT17 when evaluating learned-based tracking methods, before providing a comparison with state-of-the-art methods for completeness.

5.6.1 MOT17 Dataset Limitations

MOT17 consists of 7 training sequences, totaling approximately 5.9K frames (215 seconds of video), with a test set of 7 others videos whose annotations are kept private. Results must be submitted through an official evaluation server that enforces a 72-hour waiting period between submissions and a maximum of 4 submissions per method.

A critical limitation of MOT17 (and similarly, MOT20) is the **absence of a proper validation set**, which severely impedes the development and evaluation of learned MOT approaches, like End-To-End (E2E) methods or CAMEL. Popular works [84, 107, 109, 114–116] commonly create a validation split using the second half of all training sequences. However, we believe this practice is methodologically unsound, especially for learned methods, as it is prone to overfitting since both portions share the same scene characteristics and often contain the same tracked identities. This lack of proper validation set prevents meaningful ablation studies and proper model validation.

MOT17's dataset design thus **inherently favors heuristic-based methods** that require only hyperparameter optimization, over data-driven approaches that need a proper training and validation set. This bias is reflected in the benchmark's leaderboard, which is dominated by heuristic trackers. On the other hand, learned approaches typically underperform on MOT17 despite their success on other datasets. E2E methods typically rely on external training data, like the CrowdHuman [147] dataset, rendering a fair comparison even more difficult.

Method	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑
End-to-End MOT					
TrackFormer [132]	-	-	-	60.0	63.4
MOTR [53]	57.8	55.7	60.3	68.6	73.4
MeMOTR [134]	58.8	58.4	59.6	71.5	72.8
MOTIP [116]	59.2	56.9	62.0	71.2	75.5
MOTRv2 [115]	62.0	60.6	**63.8**	75.0	78.6
Heuristic Association					
CenterTrack [148]	52.2	-	-	64.7	67.8
FairMOT [101]	59.3	-	-	72.3	73.7
OC-SORT [109]	61.7	-	-	76.2	76.0
ByteTrack [107]	**62.8**	-	-	**77.1**	**78.7**
GHOST [84]	**62.8**	-	-	**77.1**	**78.7**
Hybrid Association					
TADN [129]	-	-	-	60.8	69.0
TransMOT [114]	-	-	-	76.3	76.4
Learned Association					
CAMELTrack	62.4	**61.4**	63.6	76.5	78.5

Table 5.6 Comparison on the MOT17 [44] test set with the private detection setting. Only **fully online methods** are reported here for fairness. Unfortunately, most state-of-the-art SORT-based methods (BoT-SORT, Diff-MOT, Hybrid-SORT, StrongSORT, Deep OC-SORT, ...) report their performance using offline post-processing mechanisms to recover missing detections (e.g. linear interpolation).

5.6.2 Comparison with state-of-the-art methods

Results on the MOT17 test set are reported in Tab. 5.6. Despite the limitations discussed above, CAMEL outperforms all existing learned approaches and achieves competitive performance with state-of-the-art heuristic methods. A complete discussion is provided in the supplementary materials of the original paper.

6

Conclusion and Perspective

This thesis tackled three key challenges in person re-identification and tracking: occlusion handling, multi-person ambiguity, and learned association. Our solutions - BPBReID, KPR, and CAMELTrack - advanced the state-of-the-art but also opened many new research directions. In this chapter, I first summarize our contributions in light of the research questions outlined in the introduction. I then discuss their limitations and future short-term research directions. I finally share my opinionated vision for the future of re-identification and multi-object tracking.

6.1 Results Summary

Our research has yielded three key contributions to the fields of person re-identification and multi-object tracking:

BPBReiD [1] (Chapter 3) establishes a strong baseline for part-based occluded person re-identification, significantly surpassing previous work on occluded re-identification. BPBReiD has had great impact on the ReID community, ranking among the top 10 most cited person re-identification papers since 2022 according to Semantic Scholar[1], and being adopted as a baseline in many subsequent research works [2, 4, 5, 149, 150]. Our contributions are twofold:

[1] On `https://www.semanticscholar.org/`, ranking "person re-identification" paper by citation count since 2022

- We introduce a novel architecture for generating part-based ReID embeddings. Unlike previous approaches that relied on fixed feature map partitioning (which is vulnerable to imprecise part localization and could miss important appearance details) or learned attention with part discovery (which yielded suboptimal partitioning due to the lack of prior knowledge about human body topology), our method adopts a hybrid approach by learning an attention mechanism supervised with pseudo human parsing labels.
- We identify a critical limitation in the training loss used by previous part-based ReID works, which overlooked the fact that part-based embeddings are not as discriminative as global embeddings. We propose *GiLt*, a novel loss specifically designed for training part-based ReID methods that combines a cross-entropy loss on global features with a part-averaged triplet loss on local features, making it resilient to partial body occlusions and non-discriminative local appearance.

KPR [2] (Chapter 4) builds upon BPBReiD and draws inspiration from Segment Anything [86] to address multi-person occlusion scenarios. Using different keypoint prompts, KPR can selectively identify either the occluded person or the occluder within the same bounding box. Our experiments demonstrate that complementing traditional crop-based ReID with keypoint prompts creates stronger synergy between both pose estimation and ReID modules, leading to significant performance improvements in both person retrieval and multi-person pose tracking. Our work makes three key contributions:

- We are the first to identify and address the *Multi-Person Ambiguity* (MPA) problem in bounding box-based ReID, which occurs when multiple people are visible within a single input crop. We demonstrate that MPA significantly impacts state-of-the-art occluded ReID methods like BPBReiD [1].
- We propose a novel promptable part-based ReID architecture and develop BIPO, a data augmentation technique that teaches the model to effectively utilize input prompts for localizing and encoding the intended target person.
- We release *Occluded PoseTrack-ReID*, a new occluded ReID dataset with keypoint annotations, to facilitate further research in promptable re-identification.

CAMELTrack [3] (Chapter 5) introduces a novel online tracking-by-detection pipeline that combines off-the-shelf models for detection, re-identification, motion modeling, and pose estimation with a *fully learned*

association module. Our contributions are twofold:

- We identify a critical research opportunity in online Multi-Object Tracking (MOT): despite tracking-by-detection (TbD) remains the most effective paradigm for online MOT, surprisingly no previous work has successfully developed a fully learned association module. While E2E methods learn association implicitly, they cannot leverage existing expert models, and SORT-based methods still rely on heuristic association rules.
- We introduce CAMEL, the first fully learned, heuristic-free association module for online TbD. CAMEL offers a deliberately simple architecture with a dedicated efficient Association-Centric Training scheme. Overall, CAMELTrack outperforms both E2E and SORT-based methods on several popular MOT benchmarks.

6.2 Limitations and Future Directions

In this section, we identify some limitations of our three proposed methods, and discuss potential future research directions.

6.2.1 BPBreID

Universal Part-based Representation Learning under Partial Observations?

Current ReID models aim to learn a single global embedding where images of the same person cluster together. However, this paradigm faces an inherent paradox when dealing with partial observations. Consider three images of the same person: a full-body image (A) and two occluded variations - an upper-body-only image (B) and a lower-body-only image (C). This becomes particularly problematic when comparing B with C - they share no common visible features despite representing the same person. The global embedding approach implicitly assumes transitivity (if A=B and B=C, then A=C), but this property breaks down under partial observations, revealing a fundamental flaw in the single embedding space paradigm. This observation suggests that the traditional approach of mapping all images to a unified embedding space may be fundamentally flawed. ReID models do not learn truly identity-centric representations, but rather appearance-centric representations that serve as a proxy to model identity. These appearance-centric representations are learned to be invariant to pose, lighting, and viewpoint changes, but when faced with partial observations, expecting such models to produce consistent embeddings becomes theoretically questionable.

Part-based methods, which compare only mutually visible regions, offer a more principled solution to this ambiguity. This theoretical limitation is reflected in practice, where part-based representation learning, which compares only mutually visible regions, offer a more principled solution to this ambiguity, as demonstrated by numerous part-bsed ReID works including PCB [26], ISP [20], and BPBreID [1]. Rather than being a technical hack to improve image retrieval performance, they may represent a fundamental requirement for representation learning under partial observation constraints. Nevertheless, this apparent limitation of deep metric learning deserves further investigation. We identify two promising research directions. First, developing generic part-based architectures for representation learning that extend beyond human ReID and dynamically align comparable features while discarding missing information across object pairs. Such generic part-based methods would learn to build representations under partial observability without priors on the input object type, making them truly universal solutions. Second, moving beyond representation learning entirely by designing ReID models that directly process image pairs to compute similarity scores, thus avoiding the ambiguity of intermediate representation in an embedding space under partial observations. In this paradigm, the network would inherently learn to compare only mutually visible parts of the input image pair. This approach faces however two key limitations compared to the first one: the need to run the neural network for every possible image pair, leading to quadratic computational complexity with gallery size (versus running the network once per image in representation learning), and the challenge of quantifying the confidence in the output similarity score when comparing partially visible objects.

Group-aware Part-based Re-Identification

While part-based approaches offer a principled solution to handling occlusions, current methods typically combine local part similarities through simple averaging, treating all body parts with equal importance. This naive aggregation strategy overlooks the discriminative power of different parts within specific contexts. Some parts may carry more identity-distinctive information than others depending on the group of people being compared. A novel research direction would be group-aware part-based reID, where the importance of each body part is dynamically computed depending on the group within which the persons are compared. For instance, more emphasis should be put on comparing the feet of a given person if that person is the only one in the group wearing red shoes.

6.2.2 KPR

Prompting with DETR's Output Tokens or Segmentation Masks

KPR was motivated by the need to increase communication between upstream detection models and downstream ReID models, addressing the information bottleneck created by communicating only through bounding boxes. Recently, DETR [105] introduced a novel approach to object detection by reformulating it as a direct set prediction problem. At its core, DETR processes an image through a CNN backbone followed by a transformer that takes a fixed set of learnable query tokens as input. Each query token is transformed into exactly one detection output, with dedicated prediction heads that output the object's class, bounding box coordinates, and confidence score. Given DETR's prominence in object detection and its elegant one-token-per-detection paradigm, an immediate extension would be to build a ReID model that can be prompted with DETR's output detection tokens, using both the image crop and its corresponding token as input. This deeper integration could potentially improve MOT performance by helping KPR disambiguate multi-person scenarios thanks to the more detailed localization information encoded in the tokens, similar to how KPR prompted with keypoints improved pose tracking. Moreover, the prompting framework could be extended to work with various types of inputs beyond keypoints, such as segmentation masks or textual descriptions.

Part-based Image Retrieval

While KPR's promptable architecture could theoretically enable retrieval based on individual body parts, it has only been evaluated for global person retrieval so far. A promising direction would be to explicitly adapt the framework for part-specific queries, allowing users to select a specific body part and retrieve other persons with similar parts (e.g., clicking on a person's boots to find others with similar boots).

6.2.3 CAMELTrack

A Hybrid Tracking-by-detection and Detection-by-tracking Framework

A key limitation of CAMELTrack's tracking-by-detection paradigm is its inability to leverage temporal information for more accurate object detection. For instance, when an object is heavily but not fully occluded in a frame, a standard object detector might miss it, even though knowledge of the object's presence in previous frames could help recover this detection. This temporal reasoning capability is a key advantage of detection-by-tracking

approaches (e.g., tracking-by-regression or tracking-by-attention). A promising direction would be to extend CAMELTrack with a detection recovery module, similar to BUSCA [130]. This module would use trajectory forecasting models (e.g., Kalman Filter) to predict the position of each tracked object in the current frame, generating tentative detections. These forecasted detections would then be input to CAMEL, which would determine whether they correspond to objects missed by the detector, to hallucinations, or to duplicates of the detector output. This would effectively combine the strengths of tracking-by-detection [52] and detection-by-tracking [132] approaches.

End-to-End Latent Space Tracking with Off-the-shelf Models

CAMELTrack currently uses the YOLO [99] object detector, taking bounding box coordinates and corresponding image crops as input. Given DETR's [105] emergence as the dominant detection paradigm, and similar to KPR's limitations, CAMELTrack would benefit from directly using DETR's output detection tokens as one of its tracking cues. These tokens, learned through DETR's transformer architecture, contain rich contextual information beyond just spatial location - including appearance features, object-to-object relationships, and scene context that are lost when reduced to just bounding box coordinates. For instance, this richer representation could help disambiguate between two different targets with heavily overlapping bounding boxes. Operating directly in this latent space would allow the tracker to leverage this additional semantic information rather than suffering from the information bottleneck of working solely with output bounding boxes. However, this should complement rather than replace the use of dedicated ReID models, which produce more discriminative appearance representations thanks to their specialization on challenging person ReID datasets and their ability to process high-resolution image crops, unlike DETR which processes targets at a lower resolution.

6.3 My Opinionated View about the Future of ReID and MOT

6.3.1 Person Re-Identification

Person Re-Identification (ReID) has reached a remarkable level of maturity in recent years, with holistic ReID[2] showing clear signs of performance

[2]Standard re-identification with non-occluded person from various camera viewpoints.

saturation. As discussed in Section 2.1.2, performance on Market1501, the most widely used ReID benchmark, has plateaued[3] in recent years, with the latest marginal improvements primarily achieved through self-supervised pre-training on large-scale human datasets. Since 2019, some works have claimed to surpass human-level performance [16], suggesting that the core challenges of the traditional ReID task have been largely addressed. The dominant paradigm in state-of-the-art ReID systems remains consistent: all build upon encoders and deep metric learning, an approach that has proven highly effective and is unlikely to see dramatic shifts in the near future. Given this maturity, research focus has already begun shifting toward more challenging variants of the traditional ReID problem. These include occluded-ReID, cross-domain adaptation, promptable re-identification, group-aware ReID, and multi-modal ReID (e.g., cloth-changing ReID, sketch-based ReID, and infrared ReID). Moreover, I expect ReID to shift from a standalone research field to becoming an application domain that leverages general advances in representation learning and foundation models. Rather than requiring ReID-specific innovations like hand-crafted architectures or explicit domain-specific biases, future progress will primarily stem from general advances in image retrieval, deep metric learning, and self-supervised learning a trend already evident in how recent improvements come mainly from better pre-training and stronger backbone networks.

6.3.2 Multi-Object Tracking

In contrast to ReID, Multiple Object Tracking (MOT) remains a highly active research field with significant challenges ahead. Given the current state of research in deep learning and MOT, I discuss hereafter several promising future directions for MOT research. I first examine how end-to-end architectures are poised to dominate future MOT systems, followed by exploring the role of self-supervised learning in addressing their substantial data requirements. I then analyze the limitations of the traditional online/offline tracking dichotomy, and conclude by synthesizing current research advances into a vision of an ideal MOT pipeline for 2025.

End-to-End Learning in Latent Space

Current state-of-the-art tracking-by-detection pipelines, like SORT-based methods, have two key drawbacks. First, they heavily rely on heuristics for

[3] https://paperswithcode.com/sota/person-re-identification-on-market-1501

object association, as discussed extensively in Chapter 5 on CAMELTrack. Second, they build on bounding boxes, which create an information bottleneck for (i) subsequent object representation based on image crops (like ReID and pose estimation) and (ii) motion estimation with Kalman Filter. These limitations point towards a fundamental shift in MOT system design. Future state-of-the-art systems will likely operate entirely in learned latent spaces, similar to how current end-to-end MOT methods work (e.g. MOTR [53]). In such architectures, human-interpretable outputs like bounding boxes, poses, and segmentation masks, which are only useful for visualization and annotation, would be obtained through dedicated prediction heads on top of track queries, rather than being core intermediate representations in the tracking pipeline. This shift towards end-to-end architectures brings several advantages. Because detection and association are jointly optimized with a unified architecture, all auxiliary tasks (such as ReID, motion estimation, or depth estimation...) emerge naturally from the shared latent representation. Among current end-to-end (E2E) approaches, I identify two main architectural directions: SAM2 [151], which maintains a memory bank containing the spatial feature maps from the N most recent frames (where each feature map of size $H \times W \times D$ is produced by the image encoder), and DETR-based methods like MOTR [53], which maintain a compact object-centric memory through track tokens. The latter appears more promising for long-term tracking due to its explicit high-level representation of the tracked object through track queries, similar to how representations are built within ReID systems. The recent evolution of MOT systems is also reflected in two distinct paradigms: tracking-by-detection (TbD) and detection-by-tracking (DbT). While TbD methods dominate most current benchmarks, DbT offers a significant advantage: the ability to leverage temporal (past) information to enhance detection in the current frame, especially in cases of heavy occlusions. Looking forward, I believe future state-of-the-art MOT methods will adopt a hybrid architecture that combines the strengths of both paradigms: using temporal information for re-detection through attention mechanisms as in DbT, while maintaining a dedicated association module to handle challenging cases and long-term tracking as in TbD. The end-to-end MOTIP [116] tracker is, to our knowledge, the closest to this vision by integrating a dedicated association module alongside DETR's detection module, although it does not yet leverage temporal information for re-detection.

Large-scale Video Pre-training

While end-to-end (E2E) MOT methods offer compelling advantages, they currently lag behind traditional SORT-based methods on most benchmarks. I argue that this performance gap stems from their need for massive amounts of challenging[4] training data, combined with the prohibitively high cost of obtaining MOT annotations, which require frame-by-frame bounding boxes and track identities. To address this fundamental limitation, self-supervised pre-training appears to be indispensable for E2E MOT methods, following a similar trajectory to the remarkable success of Large Language Models (LLMs). In the language domain, self-supervised learning through next-token prediction has enabled models to develop rich contextual understanding without explicit supervision. Similarly, I believe MOT will benefit from extensive self-supervised pre-training on video data [152–154]. Moreover, I believe models pre-trained in a self-supervised way on tasks that require predicting future frames could naturally develop strong tracking capabilities. Indeed, if a model can predict how a video will evolve over the next few seconds, tracking should be trivial for it - after all, predicting future frames requires an implicit understanding of how each object moves and persists through time. The recent work DoRA [153] takes a step in this direction by showing how tracking-like behavior can emerge through self-supervised learning from just a single long video. The main challenge with self-supervised learning for MOT is that long-term re-identification of targets might be rare in open-source standard videos, and dedicated learning or dataset might be required to circumvent this issue. For instance, synthetic data generation through neural rendering might help in building complex training examples involving long-term scene re-entries, heavy occlusions, and similar-looking persons. Importantly, exposure to large-scale video data would naturally teach models about crucial visual cues for tracking that are difficult to capture with small annotated datasets. For instance, models could learn to attend to distinctive features like license plates in traffic scenes or jersey numbers in sports videos, without requiring explicit annotation of these attributes. As training data grows and models learn these rich visual representations, the need for dedicated architectures and hand-crafted heuristics will naturally disappear.

[4]Not only is video data heavy and frame-by-frame tracking annotation labor-intensive; to make matters worse, only challenging scenarios (e.g., occlusions, scene re-entries) are relevant for training the tracker, and these typically represent a very small portion of each annotated video.

Semi-online Tracking

Recent MOT methods are typically categorized as either online or offline, but this binary distinction is artificial and limiting. Future systems should adopt a unified semi-online framework controlled by a temporal buffer parameter, where online and offline tracking become special cases with buffer sizes of 1 and infinite frames respectively. Prior work like SUSHI [128] represent a promising step in this direction. Although SUSHI was designed as a hierarchical offline method, it could be adapted to perform semi-online tracking. The flexibility offered by a configurable buffer size would allow systems to balance latency, accuracy, and computational constraints based on application requirements. Despite its practical relevance, this semi-online paradigm remains largely unexplored in recent MOT research, which focuses almost exclusively on either fully online or offline methods. This gap likely stems from the disconnect between research and industry needs, as well as benchmarking constraints - current MOT metrics and benchmarks are not designed to evaluate the trade-off between latency and accuracy across different buffer sizes.

Building the Perfect MOT Pipeline in 2025

This section synthesizes the key insights from our previous discussions on end-to-end architectures, video pre-training, and semi-online tracking to outline what an ideal MOT system could look like in 2025. Readers interested in the detailed analysis behind each component should refer to the corresponding sections above. An ideal MOT system would build on three essential components: First, a backbone pre-trained on video data in a self-supervised manner, learning a video world model that captures concepts essential for tracking: motion, occlusions, object identities, depth, object deformation, pose variation, luminosity changes, scale variations, and scene dynamics. Second, an end-to-end architecture that would fine-tune this pre-trained backbone on MOT datasets, performing detection and association entirely in the latent space, with dedicated heads producing auxiliary human-interpretable outputs (segmentation, pose, boxes). Following recent advances on E2E architectures, this model could combine a MOTR-like [53] detection head that employ track-queries to re-detect object with temporal priors, and a dedicated MOTIP-like [116] association head for associating detection across frames, thus avoiding the well-documented detection-association conflict [115] that E2E methods suffers from. Finally, this architecture would support semi-online tracking, thanks to a buffer parameter that controls how many frames are processed together during

each tracking step, allowing systems to balance tracking accuracy and application latency constraints.

6.4 Final thoughts

Overall, our work has contributed three general-purpose methods (BPBReID and KPR for re-identification, CAMELTrack for tracking), a sport-specialized ReID model (PRTreID), formalized a new research topic with dataset and metrics (SN-GameState for minimap tracking), and released three sports datasets (Basketball ReID, Soccer ReID, and Soccer Jersey Number Recognition). These years of research have opened more questions than they've closed. I'm excited to see how re-identification and tracking will evolve, particularly tracking which has many years ahead due to its complexity and need for dedicated architectures. Despite infrastructure limitations and time constraints, I believe our work has brought modest but hopefully useful contributions. I'm excited to continue this research in industry, where access to more resources will enable experiments with video foundation models and heavy end-to-end MOT methods.

Note on LLM Usage

This thesis was completed between January and February 2025, with the assistance of Large Language Models (specifically Claude 3.5 Sonnet). During this period, LLMs proved to be valuable tools for non-native English speakers, enabling more efficient writing with appropriate academic vocabulary and phrasing. However, they demonstrated fundamental limitations in generating creative ideas or conducting the deep theoretical analysis required at the doctoral level. While LLMs were instrumental in refining the presentation and language of this text, they did not contribute to its conceptual foundations or theoretical arguments. The first chapter was written without LLM assistance, while the subsequent two chapters were composed following the advent of conversational AI, marked by OpenAI's release of ChatGPT in November 2022.

Bibliography

[1] V. Somers, C. De Vleeschouwer, and A. Alahi, "Body part-based representation learning for occluded person re-identification," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, (Waikoloa, HI, USA), pp. 1613–1623, Inst. Electr. Electron. Eng. (IEEE), Jan. 2023.

[2] V. Somers, A. Alahi, and C. D. Vleeschouwer, "Keypoint promptable re-identification," in *Eur. Conf. Comput. Vis. (ECCV)*, vol. 15137 of *Lect. Notes Comput. Sci.*, pp. 216–233, Springer Nat. Switz., Nov. 2024.

[3] V. Somers*, B. Standaert*, V. Joos*, C. De Vleeschouwer, and A. Alahi, "Cameltrack: Context-aware multi-cue exploitation for online multi-object tracking," Under Review, Nov. 2023.

[4] V. Somers*, V. Joos*, S. Giancola, A. Cioppa, S. A. Ghasemzadeh, F. Magera, B. Standaert, A. M. Mansourian, X. Zhou, S. Kasaei, B. Ghanem, A. Alahi, M. Van Droogenbroeck, and C. De Vleeschouwer, "SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Seattle, WA, USA), Jun. 2024.

[5] V. Somers*, A. M. Mansourian*, C. De Vleeschouwer, and S. Kasaei, "Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking," in *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, ACM, Oct. 2023.

[6] S. Giancola, A. Cioppa, A. Deliège, F. Magera, V. Somers, L. Kang, X. Zhou, O. Barnich, C. De Vleeschouwer, A. Alahi, B. Ghanem, and

e. a. Van Droogenbroeck, Marc, "SoccerNet 2022 challenges results," in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, ACM, Oct. 2022.

[7] A. Cioppa, S. Giancola, V. Somers, F. Magera, X. Zhou, H. Mkhallati, A. Deliège, J. Held, C. Hinojosa, A. M. Mansourian, P. Miralles, O. Barnich, C. De Vleeschouwer, A. Alahi, B. Ghanem, and e. a. Van Droogenbroeck, Marc, "SoccerNet 2023 challenges results," *arXiv*, vol. abs/2309.06006, 2023.

[8] G. Van Zandycke*, V. Somers*, M. Istasse*, C. D. Don, and D. Zambrano, "DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations," in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, ACM, Oct. 2022.

[9] M. Istasse*, V. Somers*, P. Elancheliyan, J. De, and D. Zambrano, "DeepSportradar-v2: A multi-sport computer vision dataset for sport understandings," in *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, ACM, Oct. 2023.

[10] V. Joos*, V. Somers*, and B. Standaert*, "TrackLab." `https://github.com/TrackingLaboratory/tracklab`, 2024.

[11] B. Kiefer, M. Kristan, J. Pers, L. Zust, F. Poiesi, V. Somers, C. D. Vleeschouwer, A. Alahi, and F. A. de Alcantara Andrade et al., "1st workshop on maritime computer vision (macvi) 2023: Challenge results," in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, pp. 265–302, IEEE, 2023.

[12] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 2872–2893, Jun. 2022.

[13] Y. Peng, S. Hou, C. Cao, X. Liu, Y. Huang, and Z. He, "Deep learning-based occluded person re-identification: A survey," *arXiv*, vol. abs/2207.14452, 2022.

[14] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1116–1124, IEEE Computer Society, 2015.

[15] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, (Long Beach, CA, USA), pp. 1487–1495, Inst. Electr. Electron. Eng. (IEEE), Jun. 2019.

[16] 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," *CoRR*, vol. abs/1711.08184, 2017.

[17] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person ReID," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Seattle, WA, USA), pp. 11741–11749, Inst. Electr. Electron. Eng. (IEEE), Jun. 2020.

[18] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *AAAI Conf. Artif. Intell.*, vol. 36, pp. 2540–2549, Association for the Advancement of Artificial Intelligence (AAAI), Jun. 2022.

[19] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Seattle, WA, USA), pp. 6448–6457, Inst. Electr. Electron. Eng. (IEEE), Jun. 2020.

[20] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Eur. Conf. Comput. Vis. (ECCV)*, vol. 12348 of *Lect. Notes Comput. Sci.*, pp. 346–363, Springer Int. Publ., 2020.

[21] J. Yang, J. Zhang, F. Yu, X. Jiang, M. Zhang, X. Sun, Y. Chen, and W.-S. Zheng, "Learning to know where to see: A visibility-aware approach for occluded person re-identification," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, (Montreal, QC, Canada), pp. 11865–11874, Inst. Electr. Electron. Eng. (IEEE), Oct. 2021.

[22] Z. Ma, Y. Zhao, and J. Li, "Pose-guided inter- and intra-part relational transformer for occluded person re-identification," in *Proceedings of the 29th ACM International Conference on Multimedia*, ACM, Oct. 2021.

[23] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer,"

in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Nashville, TN, USA), pp. 2897–2906, Inst. Electr. Electron. Eng. (IEEE), Jun. 2021.

[24] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 542–551, Inst. Electr. Electron. Eng. (IEEE), Oct. 2019.

[25] H. Huang, X. Chen, and K. Huang, "Human parsing based alignment with multi-task learning for occluded person re-identification," in *IEEE Int. Conf. Multimedia Expo (ICME)*, (London, UK), pp. 1–6, Inst. Electr. Electron. Eng. (IEEE), Jul. 2020.

[26] 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 *Eur. Conf. Comput. Vis. (ECCV)*, vol. 11208 of *Lect. Notes Comput. Sci.*, pp. 501–518, Springer Int. Publ., 2018.

[27] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, (Venice, Italy), pp. 3774–3782, Inst. Electr. Electron. Eng. (IEEE), Oct. 2017.

[28] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, (Santiago, Chile), pp. 4678–4686, Inst. Electr. Electron. Eng. (IEEE), Dec. 2015.

[29] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," *arXiv*, vol. abs/1804.02792, 2018.

[30] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Salt Lake City, UT, USA), pp. 79–88, Inst. Electr. Electron. Eng. (IEEE), Jun. 2018.

[31] Wikipedia contributors, "Social credit system — Wikipedia, the free encyclopedia," 2023. [Online; accessed 16-January-2023].

[32] W. Wang, R. A. Solovyev, A. L. Stempkovsky, D. V. Telpukhov, and A. Volkov, "Method for whale re-identification based on siamese nets and adversarial training," *Optical Memory and Neural Networks*, vol. 29, pp. 118–132, 2020.

[33] "Swingvision: A.i. scoring, stats & line calling for tennis." https://swing.tennis/. (Accessed on 01/16/2023).

[34] Wikipedia contributors, "Sportradar — Wikipedia, the free encyclopedia," 2022. [Online; accessed 16-January-2023].

[35] "Botronics."

[36] "Mavic pro - dji." https://www.dji.com/be/mavic. (Accessed on 01/16/2023).

[37] "Robots programmed to follow you - epfl." https://actu.epfl.ch/news/robots-programmed-to-follow-you-6/. (Accessed on 01/16/2023).

[38] "Avalon | uav/drone vision." https://seadronessee.cs.uni-tuebingen.de/avalon. (Accessed on 01/16/2023).

[39] M. Buzzelli, A. Albé, and G. Ciocca, "A vision-based system for monitoring elderly people at home," *Applied Sciences*, 2020.

[40] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE Int. Conf. Image Process. (ICIP)*, (Beijing, China), pp. 3645–3649, Inst. Electr. Electron. Eng. (IEEE), Sept. 2017.

[41] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti, "3d reconstruction of freely moving persons for re-identification with a depth sensor," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4512–4519, 2014.

[42] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv*, vol. abs/1703.07737, 2017.

[43] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T. Kim, "Multiple object tracking: A literature review," *Artif. Intell.*, vol. 293, p. 103448, 2021.

[44] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *CoRR*, vol. abs/1603.00831, 2016.

[45] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, and P. Luo, "DanceTrack: Multi-object tracking in uniform appearance and diverse motion," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (New Orleans, LA, USA), pp. 20961–20970, Inst. Electr. Electron. Eng. (IEEE), Jun. 2022.

[46] Y. Cui, C. Zeng, X. Zhao, Y. Yang, G. Wu, and L. Wang, "SportsMOT: A large multi-object tracking dataset in multiple sports scenes," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, (Paris, Fr.), pp. 9887–9897, Inst. Electr. Electron. Eng. (IEEE), Oct. 2023.

[47] A. Cioppa, S. Giancola, A. Deliege, L. Kang, X. Zhou, Z. Cheng, B. Ghanem, and M. Van Droogenbroeck, "Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3491–3502, 2022.

[48] A. Doering, D. Chen, S. Zhang, B. Schiele, and J. Gall, "PoseTrack21: A dataset for person search, multi-object tracking and multi-person pose tracking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (New Orleans, LA, USA), pp. 20931–20940, Inst. Electr. Electron. Eng. (IEEE), Jun. 2022.

[49] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 319–336, Feb. 2009.

[50] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set forămulti-target, multi-camera tracking," in *Computer Vision - ECCV Workshops*, pp. 17–35, Springer Int. Publ., 2016.

[51] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "HOTA: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 548–578, Oct. 2020.

[52] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 941–951, Inst. Electr. Electron. Eng. (IEEE), Oct. 2019.

[53] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "Motr: End-to-end multiple-object tracking withătransformer," in *Eur. Conf. Comput. Vis. (ECCV)*, (Cham), pp. 659–675, Springer Nat. Switz., 2022.

[54] Z. Zhang, H. Zhang, and S. Liu, "Person re-identification using heterogeneous local graph attention networks," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Nashville, TN, USA), pp. 12131–12140, Inst. Electr. Electron. Eng. (IEEE), Jun. 2021.

[55] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, (Venice, Italy), pp. 3239–3248, Inst. Electr. Electron. Eng. (IEEE), Oct. 2017.

[56] Y. Xu, L. Zhao, and F. Qin, "Dual attention-based method for occluded person re-identification," *Knowledge-Based Systems*, vol. 212, p. 106554, Jan. 2021.

[57] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, (Venice, Italy), pp. 3980–3989, Inst. Electr. Electron. Eng. (IEEE), Oct. 2017.

[58] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Honolulu, HI, USA), pp. 20–28, Inst. Electr. Electron. Eng. (IEEE), Jul. 2017.

[59] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, pp. 4500–4509, Sept. 2019.

[60] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Long Beach, CA, USA), pp. 667–676, Inst. Electr. Electron. Eng. (IEEE), Jun. 2019.

[61] H. Wang, X. Chen, and C. Liu, "Pose-guided part matching network via shrinking and reweighting for occluded person re-identification," *Image Vis. Comput.*, vol. 111, p. 104186, Jul. 2021.

[62] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," *Computer Vision - ECCV 2018*, pp. 418–437, 2018.

[63] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Salt Lake City, UT, USA), pp. 1062–1071, Inst. Electr. Electron. Eng. (IEEE), Jun. 2018.

[64] G. Qi, G. Hu, X. Wang, N. Mazur, Z. Zhu, and M. Haner, "EXAM: A framework of learning extreme and moderate embeddings for person re-ID," *J. Imaging*, vol. 7, pp. 1–16, Jan. 2021.

[65] H. Tan, X. Liu, B. Yin, and X. Li, "MHSA-net: Multihead self-attention network for occluded person re-identification," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, pp. 8210–8224, Nov. 2023.

[66] Q. Yang, P. Wang, Z. Fang, and Q. Lu, "Focus on the visible regions: Semantic-guided alignment model for occluded person re-identification," *Sensors*, vol. 20, pp. 1–15, Aug. 2020.

[67] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Long Beach, CA, USA), pp. 11969–11978, Inst. Electr. Electron. Eng. (IEEE), Jun. 2019.

[68] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Las Vegas, NV, USA), pp. 2818–2826, Inst. Electr. Electron. Eng. (IEEE), Jun. 2016.

[69] G. Adaimi, S. Kreiss, and A. Alahi, "Deep visual re-identification with confidence," *arXiv*, vol. abs/1906.04692, 2019.

[70] K. Zhu, H. Guo, S. Zhang, Y. Wang, J. Liu, J. Wang, and M. Tang, "AAformer: Auto-aligned transformer for person re-identification," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, pp. 17307–17317, Dec. 2024.

[71] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Honolulu, HI, USA), pp. 3652–3661, Inst. Electr. Electron. Eng. (IEEE), Jul. 2017.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Las Vegas, NV, USA), pp. 770–778, Inst. Electr. Electron. Eng. (IEEE), Jun. 2016.

[73] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-net," *Computer Vision - ECCV 2018*, pp. 484–500, 2018.

[74] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for general instance re-identification," *arXiv preprint arXiv:2006.02631*, 2020.

[75] X. Zang, G. Li, W. Gao, and X. Shu, "Learning to disentangle scenes for person re-identification," *Image Vis. Comput.*, vol. 116, p. 104330, Dec. 2021.

[76] C. Yan, G. Pang, J. Jiao, X. Bai, X. Feng, and C. Shen, "Occluded person re-identification with single-scale global representations," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, (Montreal, QC, Canada), pp. 11855–11864, Inst. Electr. Electron. Eng. (IEEE), Oct. 2021.

[77] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Long Beach, CA, USA), pp. 5686–5696, Inst. Electr. Electron. Eng. (IEEE), Jun. 2019.

[78] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, Apr. 2015.

[79] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI Conf. Artif. Intell.*, vol. 34, pp. 13001–13008, Association for the Advancement of Artificial Intelligence (AAAI), Apr. 2020.

[80] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q. Chen, and R. Ji, "Occlude them all: Occlusion-aware attention network for occluded person re-ID," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, (Montreal, QC, Canada), pp. 11813–11822, Inst. Electr. Electron. Eng. (IEEE), Oct. 2021.

[81] M. Kiran, R. G. Praveen, L. T. Nguyen-Meidine, S. Belharbi, L.-A. Blais-Morin, and E. Granger, "Holistic guidance for occluded person re-identification," *arXiv*, vol. abs/2104.06524, 2021.

[82] K. Zheng, C. Lan, W. Zeng, J. Liu, Z. Zhang, and Z.-J. Zha, "Pose-guided feature learning with knowledge distillation for occluded person re-identification," *arXiv*, vol. abs/2108.00139, 2021.

[83] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (New Orleans, LA, USA), pp. 4744–4753, Inst. Electr. Electron. Eng. (IEEE), Jun. 2022.

[84] J. Seidenschwarz, G. Brasó, V. C. Serrano, I. Elezi, and L. Leal-Taixé, "Simple cues lead to a strong multi-object tracker," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 13813–13823, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[85] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Tracking millions of humans in crowded spaces," *Group and Crowd Behavior for Computer Vision*, pp. 115–135, 2017.

[86] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, and et al., "Segment anything," *arXiv*, vol. abs/2304.02643, 2023.

[87] 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 *Eur. Conf. Comput. Vis. (ECCV)*, vol. 8693 of *Lect. Notes Comput. Sci.*, pp. 740–755, Springer Int. Publ., 2014.

[88] A. Doering, D. Chen, S. Zhang, B. Schiele, and J. Gall, "PoseTrack21: A dataset for person search, multi-object tracking and multi-person pose tracking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (New Orleans, LA, USA), pp. 20931–20940, Inst. Electr. Electron. Eng. (IEEE), Jun. 2022.

[89] M. Jia, X. Cheng, Y. Zhai, S. Lu, S. Ma, Y. Tian, and J. Zhang, "Matching on sets: Conquer occluded person re-identification without alignment," in *AAAI Conf. Artif. Intell.*, vol. 35, pp. 1673–1681, Association for the Advancement of Artificial Intelligence (AAAI), May 2021.

[90] Q. Liu, Z. Xu, G. Bertasius, and M. Niethammer, "SimpleClick: Interactive image segmentation with simple vision transformers," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, (Paris, Fr.), pp. 22233–22243, Inst. Electr. Electron. Eng. (IEEE), Oct. 2023.

[91] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9992–10002, IEEE, 2021.

[92] G. Zhang, Y. Zhang, T. Zhang, B. Li, and S. Pu, "PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 14133–14142, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[93] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Nashville, TN, USA), pp. 2917–2927, Inst. Electr. Electron. Eng. (IEEE), Jun. 2021.

[94] J. Luiten, A. Osuep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "HOTA: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 548–578, Oct. 2020.

[95] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv*, vol. abs/1504.01942, 2015.

[96] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun, "Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 15050–15061, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[97] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li, and D. Chen, "Unsupervised pre-training for person re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Nashville, TN, USA), pp. 14745–14754, Inst. Electr. Electron. Eng. (IEEE), Jun. 2021.

[98] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, (Montreal, QC, Canada), pp. 14993–15002, Inst. Electr. Electron. Eng. (IEEE), Oct. 2021.

[99] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," *arXiv*, vol. abs/2107.08430, 2021.

[100] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Eur. Conf. Comput. Vis. (ECCV)*, vol. 12356 of *Lect. Notes Comput. Sci.*, pp. 107–122, Springer Int. Publ., 2020.

[101] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 3069–3087, Sept. 2021.

[102] A. Cioppa, S. Giancola, A. Deliege, L. Kang, X. Zhou, Z. Cheng, B. Ghanem, and M. Van Droogenbroeck, "SoccerNet-tracking: Multiple object tracking dataset and benchmark in soccer videos," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, (New Orleans, LA, USA), pp. 3490–3501, Inst. Electr. Electron. Eng. (IEEE), Jun. 2022.

[103] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Seattle, WA, USA), pp. 2633–2642, Inst. Electr. Electron. Eng. (IEEE), Jun. 2020.

[104] S. Ding, L. Schneider, M. Cordts, and J. Gall, "ADA-track: End-to-end multi-camera 3D multi-object tracking with alternating detection and association," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Seattle, WA, USA), pp. 15184–15194, Inst. Electr. Electron. Eng. (IEEE), Jun. 2024.

[105] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis. (ECCV)*, vol. 12346, pp. 213–229, Springer Nat. Switz., 2020.

[106] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *IEEE Int. Conf. Image Process. (ICIP)*, (Phoenix, AZ, USA), pp. 3464–3468, Inst. Electr. Electron. Eng. (IEEE), Sept. 2016.

[107] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," in *Eur. Conf. Comput. Vis. (ECCV)*, vol. 13682 of *Lect. Notes Comput. Sci.*, pp. 1–21, Springer Nat. Switz., 2022.

[108] W. Lv, Y. Huang, N. Zhang, R.-S. Lin, M. Han, and D. Zeng, "DiffMOT: A real-time diffusion-based multiple object tracker with non-linear prediction," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Seattle, WA, USA), pp. 19321–19330, Inst. Electr. Electron. Eng. (IEEE), Jun. 2024.

[109] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 9686–9696, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[110] M. Yang, G. Han, B. Yan, W. Zhang, J. Qi, H. Lu, and D. Wang, "Hybrid-SORT: Weak cues matter for online multi-object tracking," in *AAAI Conf. Artif. Intell.*, vol. 38, pp. 6504–6512, Association for the Advancement of Artificial Intelligence (AAAI), Mar. 2024.

[111] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-SORT: Robust associations multi-pedestrian tracking," *arXiv*, vol. abs/2206.14651, 2022.

[112] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "StrongSORT: Make DeepSORT great again," *IEEE Trans. Multimedia*, vol. 25, pp. 8725–8737, 2023.

[113] G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani, "Deep OC-sort: Multi-pedestrian tracking by adaptive re-identification," in *IEEE Int. Conf. Image Process. (ICIP)*, (Kuala Lumpur, Malaysia), pp. 3025–3029, Inst. Electr. Electron. Eng. (IEEE), Oct. 2023.

[114] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "TransMOT: Spatial-temporal graph transformer for multiple object tracking," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, (Waikoloa, HI, USA), pp. 4859–4869, Inst. Electr. Electron. Eng. (IEEE), Jan. 2023.

[115] Y. Zhang, T. Wang, and X. Zhang, "MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 22056–22065, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[116] R. Gao, Y. Zhang, and L. Wang, "Multiple object tracking as ID prediction," *arXiv*, vol. abs/2403.16848, 2024.

[117] Z. Qin, S. Zhou, L. Wang, J. Duan, G. Hua, and W. Tang, "MotionTrack: Learning robust short-term and long-term motions for multi-object tracking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 17939–17948, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[118] X. Han, N. Oishi, Y. Tian, E. Ucurum, R. Young, C. Chatwin, and P. Birch, "ETTrack: Enhanced temporal motion predictor for multi-object tracking," *arXiv*, vol. abs/2405.15755, 2024.

[119] M. Adžemović, P. Tadić, A. Petrović, and M. Nikolić, "Engineering an efficient object tracker for non-linear motion," *arXiv*, vol. abs/2407.00738, 2024.

[120] M. Adžemović, P. Tadić, A. Petrović, and M. Nikolić, "Beyond kalman filters: Deep learning-based filters for improved object tracking," *arXiv*, vol. abs/2402.09865, 2024.

[121] C. Xiao, Q. Cao, Z. Luo, and L. Lan, "MambaTrack: A simple baseline for multiple object tracking with state space model," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4082–4091, ACM, Oct. 2024.

[122] H.-W. Huang, C.-Y. Yang, W. Chai, Z. Jiang, and J.-N. Hwang, "Exploring learning-based motion models in multi-object tracking," *arXiv*, vol. abs/2403.10826, 2024.

[123] Y.-H. Wang, J.-W. Hsieh, P.-Y. Chen, M.-C. Chang, H.-H. So, and X. Li, "SMILEtrack: SiMIlarity LEarning for occlusion-aware multiple object tracking," in *AAAI Conf. Artif. Intell.*, vol. 38, pp. 5740–5748, Association for the Advancement of Artificial Intelligence (AAAI), Mar. 2024.

[124] H. Ren, S. Han, H. Ding, Z. Zhang, H. Wang, and F. Wang, "Focus on details: Online multi-object tracking with diverse fine-grained representation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 11289–11298, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[125] H. Hashempoor, R. Koikara, and Y. D. Hwang, "FeatureSORT: Essential features for effective tracking," *arXiv*, vol. abs/2407.04249, 2024.

[126] G. Braso and L. Leal-Taixe, "Learning a neural solver for multiple object tracking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Seattle, WA, USA), pp. 6246–6256, Inst. Electr. Electron. Eng. (IEEE), Jun. 2020.

[127] X. Zhou, T. Yin, V. Koltun, and P. Krahenbuhl, "Global tracking transformers," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (New Orleans, LA, USA), pp. 8761–8770, Inst. Electr. Electron. Eng. (IEEE), Jun. 2022.

[128] O. Cetintas, G. Brasó, and L. Leal-Taixé, "Unifying short and long-term tracking with graph hierarchies," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 22877–22887, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[129] A. Psalta, V. Tsironis, and K. Karantzalos, "Transformer-based assignment decision network for multiple object tracking," *Comput. Vis. Image Underst.*, vol. 241, p. 103957, Apr. 2024.

[130] L. Vaquero, Y. Xu, X. Alameda-Pineda, V. M. Brea, and M. Mucientes, "Lost and found: Overcoming detector failures in online multi-object tracking," *arXiv*, vol. abs/2407.10151, 2024.

[131] J. Xu, Y. Cao, Z. Zhang, and H. Hu, "Spatial-temporal relation networks for multi-object tracking," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 3987–3997, Inst. Electr. Electron. Eng. (IEEE), Oct. 2019.

[132] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: Multi-object tracking with transformers," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (New Orleans, LA, USA), pp. 8834–8844, Inst. Electr. Electron. Eng. (IEEE), Jun. 2022.

[133] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: Multiple object tracking with transformer," *arXiv*, vol. abs/2012.15460, 2020.

[134] R. Gao and L. Wang, "MeMOTR: Long-term memory-augmented transformer for multi-object tracking," in *IEEE/CVF Int. Conf. Comput.*

Vis. (ICCV), (Paris, Fr.), pp. 9867–9876, Inst. Electr. Electron. Eng. (IEEE), Oct. 2023.

[135] F. Yan, W. Luo, Y. Zhong, Y. Gan, and L. Ma, "Bridging the gap between end-to-end and non-end-to-end multi-object tracking," *arXiv*, vol. abs/2305.12724, 2023.

[136] S. Chen, E. Yu, J. Li, and W. Tao, "Delving into the trajectory long-tail distribution for muti-object tracking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Seattle, WA, USA), pp. 19341–19351, Inst. Electr. Electron. Eng. (IEEE), Jun. 2024.

[137] S. You, H. Yao, B.-k. Bao, and C. Xu, "UTM: A unified multiple object tracking model with identity-aware feature enhancement," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Vancouver, Can.), pp. 21876–21886, Inst. Electr. Electron. Eng. (IEEE), Jun. 2023.

[138] J. Cai, M. Xu, W. Li, Y. Xiong, W. Xia, Z. Tu, and S. Soatto, "MeMOT: Multi-object tracking with memory," *arXiv*, vol. abs/2203.16761, 2022.

[139] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, Jun. 2019.

[140] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, vol. abs/1807.03748, 2018.

[141] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-time multi-person pose estimation based on MMPose," *arXiv*, vol. abs/2303.07399, 2023.

[142] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution vision transformer for dense predict," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, eds.), pp. 7281–7293, 2021.

[143] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv*, vol. abs/1711.05101, 2017.

[144] A. Doering and J. Gall, "A gated attention transformer for multi-person pose tracking," in *IEEE/CVF Int. Conf. Comput. Vis. Work. (ICCV Work.)*, (Paris, Fr.), pp. 3181–3190, Inst. Electr. Electron. Eng. (IEEE), Oct. 2023.

[145] H.-W. Huang, C.-Y. Yang, J. Sun, P.-K. Kim, K.-J. Kim, K. Lee, C.-I. Huang, and J.-N. Hwang, "Iterative scale-up ExpansionIoU and deep features association for multi-object tracking in sports," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. Work. (WACVW)*, (Waikoloa, HI, USA), pp. 163–172, Inst. Electr. Electron. Eng. (IEEE), Jan. 2024.

[146] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[147] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *CoRR*, vol. abs/1805.00123, 2018.

[148] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Eur. Conf. Comput. Vis. (ECCV)* (A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds.), vol. 12349, pp. 474–490, Springer, 2020.

[149] Y. Zhang, Y. Yang, W. Kang, and J. Zhen, "Multi-scale occlusion suppression network for occluded person re-identification," *Pattern Recognit. Lett.*, vol. 185, pp. 66–72, 2024.

[150] S. Dou, X. Jiang, Y. Tu, J. Gao, Z. Qu, Q. Zhao, and C. Zhao, "DROP: decouple re-identification and human parsing with task-specific features for occluded person re-identification," *CoRR*, vol. abs/2401.18032, 2024.

[151] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C. Wu, R. B. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," *CoRR*, vol. abs/2408.00714, 2024.

[152] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. G. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," *CoRR*, vol. abs/2404.08471, 2024.

[153] S. Venkataramanan, M. N. Rizve, J. Carreira, Y. M. Asano, and Y. Avrithis, "Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024.

[154] Z. Lu, B. Shuai, Y. Chen, Z. Xu, and D. Modolo, "Self-supervised multi-object tracking with path consistency," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 19016–19026, IEEE, 2024.