



Delft University of Technology

Document Version

Final published version

Citation (APA)

Krebs, S. A. (2026). *Visual tracking of vulnerable road users for automated driving*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:7d6a1a30-fee8-4730-82c3-8bc8f532f279>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

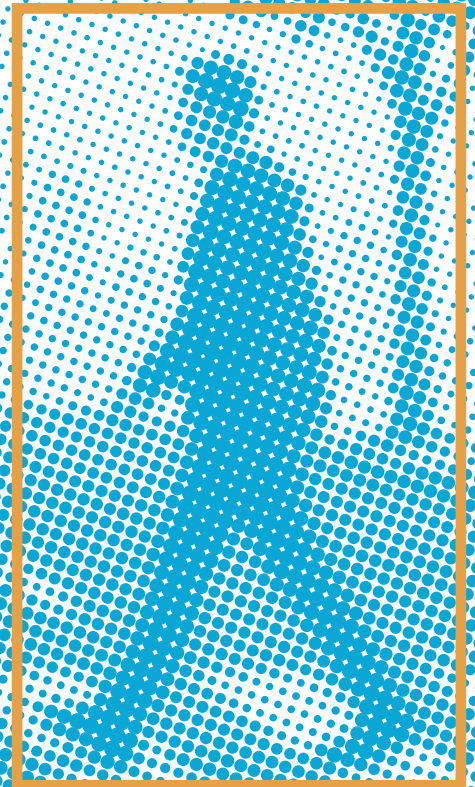
Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

VISUAL TRACKING OF VULNERABLE ROAD USERS FOR AUTOMATED DRIVING

Sebastian Krebs



VISUAL TRACKING OF VULNERABLE ROAD USERS FOR AUTOMATED DRIVING

VISUAL TRACKING OF VULNERABLE ROAD USERS FOR AUTOMATED DRIVING

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
Prof.dr.ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Friday, 10 July 2026, 12:30

by

Sebastian Alexander KREBS

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. D.M. Gavrilă	Delft University of Technology, promotor
Dr. J.E.P. Kooij	Delft University of Technology, promotor

Independent members:

Prof. dr. ir. R. Happee	Delft University of Technology
Prof. dr. T. Gevers	University of Amsterdam
Prof. dr. A. Valada	University of Freiburg, Germany
Dr. H. Caesar	Delft University of Technology
Dr. J.C. van Gemert	Delft University of Technology
Prof. dr. ir. J.C.F. de Winter	Delft University of Technology, reserve member



Keywords: Multi-Object Tracking, Vulnerable Road Users, Tracking Datasets, Person Re-Identification, Automated Driving

Cover by: Kim V. Klamser

Copyright © 2026 by S. Krebs

ISBN 978-94-6518-339-8

An electronic version of this dissertation is available at
<http://repository.tudelft.nl/>.

Without data, you're just another person with an opinion.

W. Edwards Deming

CONTENTS

Acronyms	xi
Summary	xiii
Samenvatting	xv
1 Introduction	1
1.1 Motivation, Scope, Challenges and Research Questions	1
1.1.1 Road Safety and Vulnerable Road Users	2
1.1.2 From Driver Assistance to Fully Autonomous Driving	5
1.1.3 Scope and Positioning of this Thesis	7
1.1.4 Key Challenges for Visual Tracking of VRUs	8
1.1.5 Research Questions of this Thesis	13
1.2 Thesis Outline and Contributions	13
1.2.1 Previous Work: Tracking with Deep Learning	14
1.2.2 Semi-supervised Creation of an AD Person Dataset	15
1.2.3 ECP2.0 Dataset for Person Tracking in AD	16
1.2.4 Learning Appearance Embeddings for Person ReID	16
2 Previous Work	19
2.1 Traditional Object Tracking	19
2.2 Tracking with Deep Learning	21
2.2.1 Deep Learning for Tracking: Until 2017	21
2.2.2 Deep Learning for Tracking: Since 2017	26
2.3 Dataset Generation	30
2.4 Tracking Datasets in AD	33
2.5 Person Re-Identification	34
2.6 Discussion	35
3 Semi-supervised Pipeline for Creating a Person Tracking Dataset	37
3.1 Pipeline Overview	37
3.2 Data Preparation and Preprocessing	39
3.2.1 Detection and Appearance Embeddings Generation	39
3.2.2 3D Uplifting	39
3.2.3 Ego-Motion Compensation	40
3.3 Graph-based Track Creation	41
3.4 Tracklet-based Track Creation	43
3.4.1 Tracklet Generation	43
3.4.2 Tracklet Merging	44

3.5	Batch Aggregation and Temporal Refinement	46
3.6	Experimental Validation	46
3.6.1	Comparison of Pseudo GT Generation Methods	47
3.6.2	Ablation Studies	48
3.6.3	Ceiling Study (Perfect Detector)	49
3.7	Discussion	51
	Appendices	53
3.A	Appearance Embedding Network	53
3.B	Additional Qualitative Results	53
4	The ECP2.0 Dataset for Person Tracking in AD	55
4.1	Data Recordings	55
4.2	Annotations	56
4.2.1	Manual Annotations	56
4.2.2	Pseudo Annotations	57
4.2.3	On the Use of 2D Bounding Boxes	57
4.3	Dataset Statistics	57
4.4	Benchmarking	59
4.5	Tracking Evaluation Metrics	59
4.6	Validation Dataset	60
4.7	Evaluation of pseudo GT	60
4.7.1	Keyframe Frequency	61
4.7.2	Pseudo Ground-Truth - Localization Accuracy	61
4.7.3	Dependence on Object Height	62
4.8	Generalization Capabilities	63
4.9	Use of Pseudo Ground-Truth for Benchmarking	63
4.10	Discussion	65
	Appendices	67
4.A	Temporal Correspondence Annotation Tool	67
4.B	Hyper-Parameter Setting	67
4.C	Evaluation per Object Density	68
5	Learning Appearance Embeddings for Person Re-Identification	71
5.1	Overview	71
5.2	Proposed Approach	72
5.2.1	Camera-Only Baseline	72
5.2.2	Adding Object Point Information as Embedding	73
5.2.3	Extension to Multi-Modal Input	74
5.2.4	Auxiliary Tasks	75
5.3	Experiments	76
5.3.1	Dataset	76
5.3.2	Evaluation Metrics	76
5.3.3	Experimental Setup	77
5.3.4	Comparison to Baselines and State-of-the-Art	77
5.3.5	Impact of different Multi-Modal Fusion Strategies	78
5.3.6	Ablation Study: Auxiliary Tasks and 3D-CIPE	79

5.4	Discussion	80
6	Conclusion and Future Work	83
6.1	Conclusion	83
6.2	Future Work.	86
	Acknowledgements	93
	Curriculum Vitæ	95
	List of Publications	97
	Bibliography	99

ACRONYMS

AD	Automated Driving	37
ADAS	Advanced Driver-Assistance System	6
AEB	Autonomous Emergency Braking	3
AssA	Association Accuracy	59
AssPr	Association Precision	60
AssRe	Association Recall	60
BEV	Birds-Eye View	57
CMC	Cumulative Matching Characteristics	76
CV	Constant Velocity	46
DAG	Directed Acyclic Graph	38
DetA	Detection Accuracy	59
FoV	Field of View	55
FPC	Forest Path Cutting	44
fps	Frames per Second	25
GT	Ground-Truth	86
HOTA	Higher Order Tracking Accuracy	84
IMU	Inertial Measurement Unit	8
IoU	Intersection-over-Union	84
IP	Integer Programming	43
mAP	mean Average Precision	85
MOT	Multi-Object Tracking	27
NMS	Non-Maximum Suppression	27
NHTSA	U.S. National Highway Traffic Safety Administration	3
LocA	Localization Accuracy	60
POM	Probabilistic Occupancy Grid Map	42
ReID	Re-Identification	85
RQ	Research Question	83
RTS	Rauch-Tung-Striebel	45
VRU	Vulnerable Road User	85
WHO	World Health Organization	2

SUMMARY

This thesis addresses the problem of vision-based tracking of Vulnerable Road Users (VRUs) such as pedestrians and riders in the context of Automated Driving (AD). The work primarily builds on camera-based perception, while also considering LiDAR as a complementary sensor. Robust tracking of persons is a critical perception task, as it enables downstream modules for motion prediction and safe planning. The motivation stems from the disproportionate involvement of VRUs in traffic fatalities, as globally pedestrians and riders account for more than half of all road deaths, particularly in urban environments. By providing temporally consistent trajectories and identity preservation across occlusion and re-entries, tracking contributes to safer and more anticipatory automated driving systems.

The thesis first covers an extended review of the state-of-the-art in multi-object tracking. A modular framework is outlined to describe traditional tracking pipelines. The review then focuses on deep learning-based methods. Early methods typically extended classical pipelines with deep features, while later works moved toward end-to-end formulations and transformer-based architectures. This survey provides the methodological foundation and highlights the transition toward data-driven approaches.

As these approaches rely on large amounts of data for both training and evaluation, fully manual annotation becomes increasingly challenging and costly, as the required effort grows rapidly with dataset size. To address this challenge, the thesis develops a semi-supervised trajectory generation pipeline. The approach leverages camera detections, LiDAR-based estimation of 3D locations from 2D boxes, and vehicle inertial sensing for ego-motion compensation. Based on this data, the approach generates temporally dense pseudo ground-truth (i.e. 2D bounding boxes, 3D person locations) from sparse, manual annotations at keyframes. Two complementary strategies are employed to generate tracks within each batch, whereby each batch is defined to begin and end at a keyframe. The first is a graph-based optimization for global association that leverages a spatio-temporal graph. The second is a tracklet-based approach that combines tracklet generation with subsequent tracklet merging. For both strategies, the outputs obtained within each batch are subsequently aggregated into complete trajectories, guided by the keyframe annotations. An experimental evaluation of these approaches, including comparisons to baselines, ablation studies, and ceiling analyses, demonstrates that the tracklet-based pipeline achieves the most robust overall performance. This approach significantly reduces annotation effort while producing temporally consistent trajectories suitable for training deep tracking models.

The EuroCity Persons (ECP) 2.0 dataset is introduced as a large-scale image dataset for person detection, tracking, and prediction in traffic. The dataset was collected on-board a vehicle driving through 29 cities in 11 European countries. It contains more than 220K unique person trajectories, in more than 2.0M images and comes with a size

of 11 TB. ECP2.0 is about one order of magnitude larger than previous state-of-the-art person datasets in automotive context. It offers remarkable diversity in terms of geographical coverage, time of day, weather and seasons. The pipeline proposed in this thesis is used to generate the dense annotations, providing temporally consistent pseudo ground-truth on top of the recorded image data. Experiments analyze the resulting annotation quality and object-dependent effects. The benefits of using the pseudo ground-truth for training and testing of state-of-the-art tracking methods are further demonstrated, showing improved tracking accuracy and robustness compared to training on sparsely annotated or purely detector-based supervision, while also highlighting remaining challenges in handling long-term occlusions and identity switches. The ECP2.0 dataset is made freely available for non-commercial research use.

Finally, this thesis improves person appearance embeddings as a means to strengthen identity association in multi-object tracking. To this end, a novel method for creating appearance embeddings to identify individual persons using an object Re-Identification ([ReID](#)) framework is introduced. CLFormer (Camera LiDAR Transformer) is presented, a transformer-based architecture that incorporates multi-modal data from both camera and LiDAR sensors. The 3D Cuboid-Inclusive Point Embedding (3D-CIPE) is introduced, which leverages rich data from LiDAR point clouds and 3D cuboids to add a learnable embedding into the transformer structure. Additionally, through ablation studies, various strategies for the early and late fusion of multi-modal input data are explored and analyzed. To evaluate the proposed CLFormer, the nuScenes dataset [14] is reinterpreted for [ReID](#) purposes and used for the experiments. CLFormer demonstrates a significant improvement in performance, outperforming the image-only baseline with an increase of 2.3 in mean Average Precision ([mAP](#)).

In conclusion, this thesis advances the state-of-the-art in [VRU](#) tracking for automated driving by providing an extended methodological review, novel semi-supervised dataset creation techniques, a large-scale benchmark dataset, and a complementary [ReID](#) method that provides stronger identity cues for tracking. Together, these contributions offer new methodological insights, practical tools, and benchmarks that can support future research and development in the field of automated driving perception and tracking. Future work may focus on further improving the semi-supervised annotation method, particularly by reducing or even moving beyond the reliance on manual keyframe annotations, toward full unsupervised or self-supervised methods. In general, the long-term vision for multi-object tracking in automated driving points toward systems that unify perception, prediction, and decision-making in end-to-end frameworks, enabling robust and reliable operation in diverse and challenging real-world environments.

SAMENVATTING

Deze dissertatie behandelt het probleem van vision-based tracking van Vulnerable Road Users (VRUs), zoals voetgangers en fietsers, in de context van Automated Driving (Automated Driving (AD)). Het werk bouwt primair voort op camera-gebaseerde perceptie, waarbij LiDAR eveneens als complementaire sensor wordt beschouwd. Robuuste tracking van personen is een kritische perceptietaak, aangezien deze downstream modules voor motion prediction en veilige planning mogelijk maakt. De motivatie komt voort uit de disproportionele betrokkenheid van kwetsbare verkeersdeelnemers (VRUs) bij verkeersdoden; wereldwijd zijn voetgangers en fietsers verantwoordelijk voor meer dan de helft van alle verkeersdoden, met name in stedelijke omgevingen. Door het leveren van temporeel consistente trajecten en het behouden van identiteit over occlusies en herintredingen heen, draagt tracking bij aan veiligere en meer anticiperende Automated Driving-systemen.

De dissertatie begint met een uitgebreide review van de state-of-the-art in multi-object tracking. Er wordt een modulair framework geschetst om traditionele tracking pipelines te beschrijven. Vervolgens richt de review zich op deep learning-gebaseerde methoden. Vroege methoden breidden doorgaans klassieke pipelines uit met deep features, terwijl latere werken evolueerden naar end-to-end formuleringen en transformer-gebaseerde architecturen. Deze survey vormt de methodologische basis en benadrukt de transitie naar data-gedreven benaderingen.

Aangezien deze benaderingen afhankelijk zijn van grote hoeveelheden data voor zowel training als evaluatie, wordt volledige handmatige annotatie steeds uitdagender en kostbaarder, omdat de benodigde inspanning snel toeneemt met de datasetgrootte. Om deze uitdaging aan te pakken, ontwikkelt deze dissertatie een semi-supervised trajectory generation pipeline. De benadering maakt gebruik van camera-detecties, LiDAR-gebaseerde schatting van 3D-locaties vanuit 2D bounding boxes, en voertuig-inertiële metingen voor ego-motion compensatie. Op basis van deze data genereert de methode temporeel dichte pseudo ground-truth (d.w.z. 2D bounding boxes en 3D persoonslocaties) vanuit sparse, handmatige annotaties op keyframes. Twee complementaire strategieën worden toegepast om tracks binnen elke batch te genereren, waarbij een batch wordt gedefinieerd als beginnend en eindigend op een keyframe. De eerste strategie is een graph-based optimalisatie voor globale associatie die gebruikmaakt van een spatio-temporele graaf. De tweede is een tracklet-based benadering die tracklet-generatie combineert met daaropvolgende tracklet-merging. Voor beide strategieën worden de binnen elke batch verkregen outputs vervolgens samengevoegd tot volledige trajecten, geleid door de keyframe-annotaties. Een experimentele evaluatie van deze benaderingen, inclusief vergelijkingen met baselines, ablation studies en ceiling-analyses, toont aan dat de tracklet-based pipeline de meest robuuste algemene prestaties behaalt. Deze aanpak reduceert de annotatie-inspanning aanzienlijk, terwijl temporeel consistente trajecten worden geproduceerd die geschikt zijn voor het trainen van deep tracking-modellen.

De EuroCity Persons (ECP) 2.0 dataset wordt geïntroduceerd als een grootschalige image dataset voor person detection, tracking en prediction in verkeersscenario's. De dataset werd verzameld aan boord van een voertuig dat reed door 29 steden in 11 Europese landen. De dataset bevat meer dan 220K unieke persoons-trajecten in meer dan 2.0M afbeeldingen en heeft een omvang van 11 TB. ECP2.0 is ongeveer een orde van grootte groter dan eerdere state-of-the-art person datasets in de automotive context. De dataset biedt een opmerkelijke diversiteit wat betreft geografische dekking, tijdstip van de dag, weersomstandigheden en seizoenen. De in deze dissertatie voorgestelde pipeline wordt gebruikt om de dichte annotaties te genereren en levert temporeel consistente pseudo ground-truth bovenop de opgenomen beelddata. Experimenten analyseren de resulterende annotatiekwaliteit en object-afhankelijke effecten. De voordelen van het gebruik van pseudo ground-truth voor training en testing van state-of-the-art tracking-methoden worden verder aangetoond, met verbeterde tracking accuracy en robuustheid in vergelijking met training op sparse annotaties of puur detector-gebaseerde supervisie, terwijl tevens resterende uitdagingen bij het omgaan met langdurige occlusies en identity switches worden belicht. De ECP2.0 dataset wordt vrij beschikbaar gesteld voor niet-commercieel wetenschappelijk onderzoek.

Ten slotte verbetert deze dissertatie person appearance embeddings als middel om identity association in multi-object tracking te versterken. Hiertoe wordt een nieuwe methode geïntroduceerd voor het creëren van appearance embeddings om individuele personen te identificeren met behulp van een object Re-Identification (ReID) framework. CLFormer (Camera LiDAR Transformer) wordt gepresenteerd als een transformer-gebaseerde architectuur die multi-modale data van zowel camera- als LiDAR-sensoren integreert. De 3D Cuboid-Inclusive Point Embedding (3D-CIPE) wordt geïntroduceerd en benut rijke informatie uit LiDAR point clouds en 3D cuboids om een leerbare embedding in de transformer-structuur te integreren. Daarnaast worden via ablation studies verschillende strategieën voor vroege en late fusie van multi-modale inputdata onderzocht en geanalyseerd. Voor de evaluatie van de voorgestelde CLFormer wordt de nuScenes dataset [14] hergeïnterpreteerd voor ReID-doeleinden en gebruikt voor experimenten. CLFormer laat een significante prestatieverbetering zien en overtreft de image-only baseline met een toename van 2.3 in mean Average Precision (mAP).

Concluderend bevordert deze dissertatie de state-of-the-art in VRU tracking voor Automated Driving door een uitgebreide methodologische review, nieuwe semi-supervised datasetcreatietechnieken, een grootschalige benchmark dataset en een complementaire ReID-methode die sterkere identity cues voor tracking levert. Gezamenlijk bieden deze bijdragen nieuwe methodologische inzichten, praktische tools en benchmarks die toekomstig onderzoek en ontwikkeling op het gebied van Automated Driving perceptie en tracking kunnen ondersteunen. Toekomstig werk kan zich richten op verdere verbetering van de semi-supervised annotatiemethode, met name door het verminderen van of zelfs afstappen van de afhankelijkheid van handmatige keyframe-annotaties, richting volledig unsupervised of self-supervised methoden. In algemene zin wijst de langetermijnvisie voor multi-object tracking in Automated Driving naar systemen die perceptie, prediction en besluitvorming verenigen in end-to-end frameworks, met als doel robuuste en betrouwbare werking in diverse en uitdagende real-world omgevingen.

1

INTRODUCTION

TRACKING objects visually is a fundamental capability that enables humans to make sense of a dynamic world. We naturally follow the motion of other people, vehicles, or animals across time, even when they briefly disappear from sight. In fact, human babies are known to be able to track objects even before having the ability to classify them into categories [149]. This early skill highlights that temporal continuity is one of the most basic building blocks of perception, preceding recognition and semantic understanding.

In everyday driving, for instance, a driver keeps track of pedestrians while they cross behind parked cars, or of cyclists reappearing after short occlusions. Such tracking is largely subconscious, yet it forms the basis of safe interactions in complex environments.

For automated driving systems, recreating this capability remains challenging: the implicit and highly evolved perceptual mechanisms that allow humans to track objects with ease must be replaced by explicit processing of sensor data and algorithmic reasoning to link detections over time, preserve identities, and recover from occlusions or re-entries. An example of such a situation is shown in Figure 1.1, illustrating the challenges that arise when tracking persons under occlusion and ego-motion. Reliable tracking is not only crucial for a consistent environment representation, but also for downstream tasks such as motion prediction and planning [9, 147, 148]. This thesis is situated in this context, investigating how tracking of vulnerable road users, from the ego-centric perspective of a moving vehicle using on-board sensors, can be advanced to support safer automated vehicles.

1.1. MOTIVATION, SCOPE, CHALLENGES AND RESEARCH QUESTIONS

This section introduces the motivation and scope of the thesis. It begins by highlighting the importance of vulnerable road users and explaining why they deserve particular attention in the development of safe automated vehicles (Section 1.1.1). The discussion then turns to the evolution from advanced driver assistance systems to higher levels of

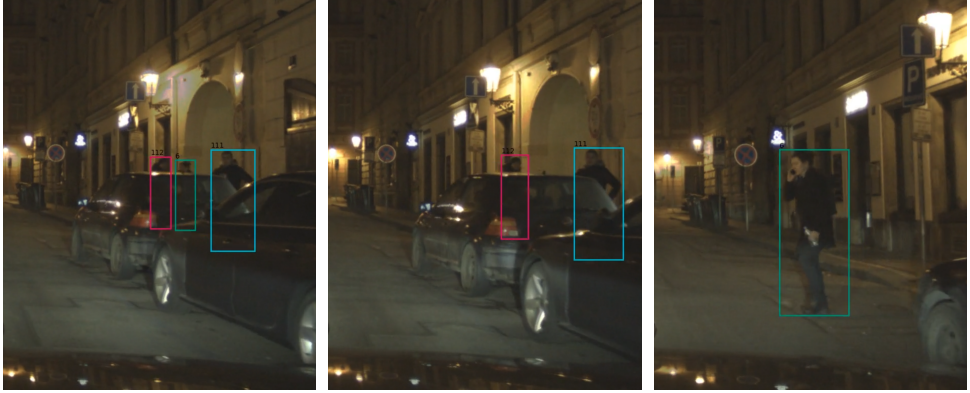


Figure 1.1: Example of visual tracking of a pedestrian from an ego-centric vehicle perspective in an urban nighttime scenario. A pedestrian is tracked across time and highlighted by a green bounding box, which indicates a consistent identity. In the first frame the person is partially occluded by the parked vehicle. As the ego vehicle moves, the pedestrian becomes fully occluded. In the final image, the pedestrian reappears closer to the ego vehicle. Other pedestrians, shown in red and blue, illustrate additional identities present in the scene. The example highlights typical challenges, including occlusion, ego-motion, and identity preservation.

driving automation, where reliable perception of the traffic scene becomes an increasingly central component (Section 1.1.2). Section 1.1.3 outlines the scope and positioning of this thesis. Section 1.1.4 summarizes the key challenges for visual tracking. Building on these challenges, Section 1.1.5 concludes by formulating the research questions that guide the contributions of this thesis.

1.1.1. ROAD SAFETY AND VULNERABLE ROAD USERS

Vulnerable Road Users (VRUs) - namely pedestrians, cyclists, and riders of two- or three-wheeled vehicles - continue to bear a disproportionate burden of global traffic fatalities. According to the 2023 World Health Organization (WHO) global status report [125], an estimated 1.19 million people lost their lives on roads in 2021. A breakdown of these fatalities by type of road user reveals that VRUs account for a significant majority. Riders of powered two- and three-wheelers represent the largest group globally at 30%, followed by pedestrians (21%) and cyclists (5%). In contrast, occupants of four-wheeled vehicles account for 25%, and other road users including bus passengers make up the remaining 19%. Figure 1.2 provides a detailed regional perspective, illustrating that these global proportions mask substantial differences across WHO regions. For instance, powered two- and three-wheelers are the dominant fatality category in South-East Asia (46%), reflecting their widespread use. In contrast, pedestrians make up the largest share of fatalities in the African Region (33%), underscoring the high risk faced by non-motorized road users, especially in regions with limited road-safety infrastructure. Meanwhile, in the European Region, 4-wheeler occupants dominate the fatality share (50%), but pedestrian (25%) and cyclist (6%) deaths remain significant. In addition to these fatalities, more than 50 million people sustain non-fatal injuries annually, posing substantial human, social, and economic costs [124]. Road traffic injuries remain a leading cause of death for children and young adults aged 5-29 years, underscoring the intergenerational

and global relevance of the problem.

In the European context, road safety figures show limited overall improvement in recent years. Preliminary data from the European Commission indicates that there were approximately 20,400 road deaths across the EU in 2023, of which about 70% in urban areas involved VRUs [42]. Despite a long-term reduction of pedestrian fatalities by 31% between 2012 and 2022, cyclist deaths have been rising in many regions [42]. Moreover, older adults are increasingly affected: seniors aged 65 or above accounted for approximately 47% of all pedestrian fatalities in 2022 [42]. These figures not only point to population-specific vulnerabilities but also reflect the increasing density and complexity of urban traffic environments. The rising numbers of seniors and cyclists involved in accidents are likely driven by heavier traffic volumes, growing competition for road space, and increased exposure of VRUs in congested city areas.

This growing density and modal diversity in city traffic environments has direct consequences for VRU safety. Urban crash data reveals that around two-thirds of serious pedestrian crashes occur while crossing roadways [43]. A detailed study of fatal pedestrian collisions in Florida, for instance, found that 51% involved a vehicle traveling straight, while another 36% occurred during vehicle turning maneuvers (left or right) at intersections or crosswalks [201]. Comparable findings are echoed in other high-income countries and point to the high-risk interaction zones between motorized and non-motorized road users. Furthermore, children, seniors, and individuals with disabilities face increased risks in these zones due to limitations in mobility, visibility, and predictability.

Another often-cited concern is visibility. Approximately 75% of fatal pedestrian crashes happen at night, frequently in poorly lit environments. While this thesis does not directly address nighttime perception, such findings nonetheless illustrate the broader limitations of both human and automated visual systems in adverse conditions. Reduced contrast between VRUs and their background, coupled with the limited time available for drivers to visually identify and respond to these low-visibility scenarios, can significantly compromise timely and accurate reaction. Real-world studies indicate that pedestrian Autonomous Emergency Braking (AEB) systems, although effective during daytime, offer negligible benefits in unlit nighttime scenarios [29, 78]. This highlights the general need for robust perception technologies that remain effective across a wide range of environmental conditions, even if the current work focuses specifically on the challenges encountered under daytime and urban traffic scenarios.

Compounding the problem are the inherent limitations of human drivers. According to the U.S. National Highway Traffic Safety Administration (NHTSA), an estimated 94% of traffic accidents are attributable to human error - stemming from distraction, fatigue, misperception, or poor decision-making [159]. In typical urban traffic, drivers are required to process a high volume of heterogeneous and rapidly changing stimuli. Human drivers frequently struggle to detect VRUs in time, particularly when they are occluded by other vehicles or emerge unexpectedly, such as a child darting out from between parked cars. Moreover, their ability to accurately estimate distance, speed, and future movement intentions is severely constrained in complex environments. Such cases highlight the limitations of human attention and reaction time in dynamically evolving urban environments.

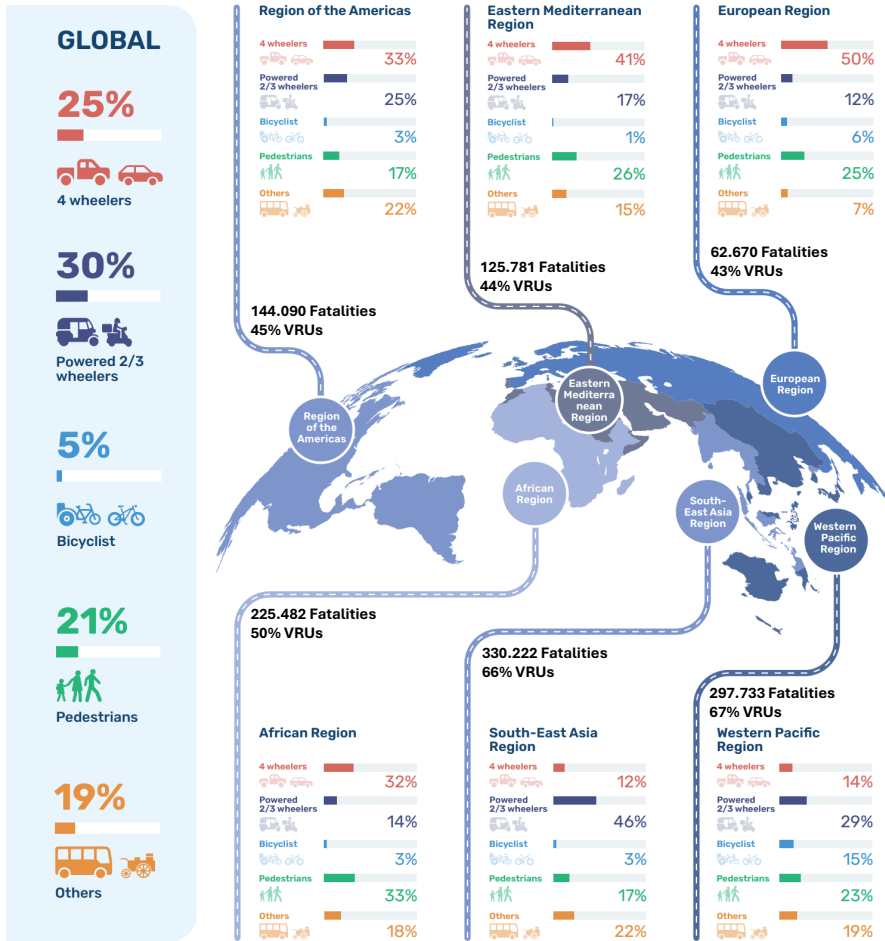


Figure 1.2: Global and regional distribution of road traffic deaths by road user type, based on the 2023 [WHO Global Status Report on Road Safety](#) [125]. Vulnerable road users (pedestrians, cyclists, and powered 2/3-wheeler riders) account for the majority of fatalities globally, with substantial regional variation. Pedestrian deaths dominate in Africa, motorcyclists in South-East Asia, and 4-wheeler occupants in Europe and the Americas. Figure adapted from [125].

In summary, the combination of high **VRU** vulnerability, complex urban environments, low-light conditions, and the limitations of human drivers strongly motivates the need for robust, automated perception systems capable of operating reliably under real-world constraints. Beyond simply detecting objects, these systems must be able to temporally associate and continuously track individual **VRUs** through occlusion, viewpoint variation, and sensor noise. Ultimately, automated perception, whether as part of advanced driver assistance systems or fully autonomous driving, is a key enabling technology for improving traffic safety outcomes and advancing long-term strategic initiatives such as the EU's Vision Zero [44] and global sustainable mobility goals.

1.1.2. FROM DRIVER ASSISTANCE TO FULLY AUTONOMOUS DRIVING

	SAE LEVEL 0™	SAE LEVEL 1™	SAE LEVEL 2™	SAE LEVEL 3™	SAE LEVEL 4™	SAE LEVEL 5™
What does the human in the driver's seat have to do?	You <u>are</u> driving whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			You <u>are not</u> driving when these automated driving features are engaged – even if you are seated in “the driver’s seat”		
	You <u>must constantly supervise</u> these support features; you must steer, brake or accelerate as needed to maintain safety			When the feature requests, you must drive	These automated driving features will not require you to take over driving	
Copyright © 2021 SAE International.						
	These are driver support features			These are automated driving features		
What do these features do?	These features are limited to providing warnings and momentary assistance	These features provide steering OR brake/acceleration support to the driver	These features provide steering AND brake/acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met	This feature can drive the vehicle under all conditions	
Example Features	• automatic emergency braking • blind spot warning • lane departure warning	• lane centering OR • adaptive cruise control	• lane centering AND • adaptive cruise control at the same time	• traffic jam chauffeur	• local driverless taxi • pedals/steering wheel may or may not be installed	• same as level 4, but feature can drive everywhere in all conditions

Figure 1.3: SAE J3016 classification of driving automation levels from 0 (no automation) to 5 (full automation). Levels 0-2 require constant human supervision, while levels 3-5 allow the system to drive under defined conditions, with increasing autonomy and reduced human involvement. Adopted from [146].

The path toward fully autonomous vehicles is best understood through the SAE J3016 framework, which categorizes vehicle automation from Level 0 (no automation) to Level 5 (full automation) [146]. These levels distinguish not only technical capability but also the role of the human driver in monitoring and controlling the vehicle. Figure 1.3 provides a detailed summary of these six levels, highlighting the transition from driver-support systems to full automation.

Levels 0 through 2 are considered driver support features. In Level 0, the driver is fully responsible, with the vehicle providing only momentary alerts or interventions, such as lane departure warnings or emergency braking. Level 1 introduces a single automated

function (e.g., adaptive cruise control), and Level 2 combines functions like lane centering and acceleration control. At all three levels, the driver is responsible for monitoring the driving environment and must be ready to intervene at any time.

Starting from Level 3, the system takes over the dynamic driving task under defined conditions. In Level 3 (conditional automation), the vehicle monitors the environment and handles driving within specific operational domains (e.g., highway traffic jam scenarios), though the human driver must remain available to take over if prompted. Level 4 (high automation) operates fully autonomously within geofenced environments, with no requirement for human takeover. Finally, Level 5 envisions a vehicle that can drive anywhere a human could, in any condition, without a driver present.

While the theoretical distinction is clear, practical deployment of current Advanced Driver-assistance Systems (ADAS) already demonstrates encouraging progress toward improved VRU safety. One of the most impactful features in this domain is Pedestrian AEB, which has consistently shown reductions in crash rates. For instance, real-world crash data indicate that Pedestrian AEB systems can reduce pedestrian crashes by approximately 27% under daylight conditions [29]. A broader study from Sweden further supports these trends, reporting crash reductions of around 8% for pedestrians and 21% for cyclists [78]. Although there remain performance limitations in nighttime conditions, both studies demonstrate a consistent trend toward improved safety outcomes. These results reflect the growing maturity and real-world relevance of such systems, even as further improvements are needed to ensure robust operation in all environments.

Given this growing evidence that AEB systems with pedestrian detection improve VRU safety - particularly under daylight conditions - their continued adoption across vehicle fleets promises broader impact. Mercedes-Benz, for example, has already equipped more than 10 million vehicles with its *Active Brake Assist with pedestrian detection* (originally introduced as *PRE-SAFE® brake with pedestrian detection*) illustrating the increasing availability and maturity of such technologies and the automotive industry's commitment to integrating pedestrian safety into mass-market vehicles. In parallel, Mercedes-Benz also introduced DRIVE PILOT in 2021, the first internationally approved SAE Level 3 automated driving system. DRIVE PILOT enables conditional automation on designated highways in daylight and favorable weather conditions up to speeds of 95 km/h (initially limited to 60 km/h at introduction in 2021). When active, it allows the driver to disengage from driving tasks, with the system assuming full responsibility for dynamic driving within its operational domain. While limitations persist, the widespread deployment of systems like pedestrian AEB and the rollout of Level 3 technologies like DRIVE PILOT represent meaningful and scalable advances toward safer and more autonomous road transport.

As automation progresses beyond Level 2, the role of perception systems becomes increasingly critical. While systems at Levels 0 through 2 primarily focus on accident prevention through selective intervention, they are optimized for low false-positive rates to avoid annoying or confusing the driver. Each activation must be highly precise and conservative, as the human driver remains the primary fallback. In contrast, for higher automation levels beginning at Level 3 the vehicle must handle all driving tasks within defined operational domains, and the human is no longer expected to constantly supervise. This elevates the cost of false negatives: a missed detection could directly lead

to unsafe behavior. Consequently, perception systems must go beyond snapshot detection and deliver full-scene understanding, including continuous tracking and robust re-identification of **VRUs** under challenging conditions such as occlusions, crowding, and unpredictable behavior. This capability is vital for safe decision-making and intent prediction, particularly in dynamic urban settings where many **VRU**-related accidents occur.

Despite the challenges, the societal benefits of advancing vehicle automation are substantial. Nearly all traffic accidents stem from human error, whether due to distraction, misjudgment, or fatigue [159], meaning that eliminating the driver from the control loop has the potential to significantly reduce crash rates and save lives. While today's Level 3 systems, such as Mercedes-Benz's DRIVE PILOT, represent an important milestone, their impact on **VRU** safety remains limited due to their restriction to highway use cases.

Before reaching full Level 4/5 automation, a parallel development has been the emergence of enhanced Level 2 systems (often referred to as *Level 2++*) which already provide partial automation in complex urban environments, including **VRU**-rich areas. While these systems do not fully remove the human driver from the loop, they offer advanced capabilities such as automated lane changes, traffic light recognition, stopping at crosswalks for pedestrians, and stop-and-go functionality. These features deliver incremental but meaningful improvements to safety in urban traffic. At the same time, their increasingly sophisticated behavior can also create a false sense of security: despite strong performance in many scenarios, the driver must remain fully responsible and attentive at all times, which can lead to mismatched expectations and over-reliance.

In parallel, recent pilot deployments of Level 4 systems in selected geofenced urban areas (e.g., by Waymo, Cruise, Baidu, Pony.ai, and others) illustrate another step toward higher autonomy. While still experimental and limited in operational scope, these efforts demonstrate the potential of automated driving in controlled urban settings and suggest that, once scaled, such systems could meaningfully improve road safety and reshape urban mobility.

Yet these developments also underscore a central point: even as Level 2++, Level 3, and early Level 4 systems emerge in parallel, all fundamentally rely on robust perception, particularly for reliably detecting, tracking, and understanding **VRUs**. Tackling these challenges remains crucial, as achieving higher levels of autonomy holds the potential to significantly advance road safety and contribute to goals like the EUs Vision Zero targets of eliminating all traffic fatalities and serious injuries [44]. Additional benefits include reductions in traffic congestion, improved fuel efficiency, lower emissions, and expanded mobility options for elderly and mobility-impaired individuals.

1.1.3. SCOPE AND POSITIONING OF THIS THESIS

This thesis investigates the problem of Multi-Object Tracking (**MOT**) and person Re-Identification (**ReID**) for **VRUs**, with a specific focus on applications in Automated Driving (**AD**). The work follows a vision-based deep learning paradigm and emphasizes image-space representations. A multi-modal sensor setup mounted on a moving vehicle is considered, consisting of camera, LiDAR and inertial measurements. While multiple sensing modalities are available, the thesis predominantly builds on visual information.

The core objective is to develop methods that operate effectively in dense, complex urban environments from an ego-centric, on-board vehicle perspective. The scope is explicitly limited to the perception layer of the automated driving pipeline and excludes downstream tasks such as motion prediction, path planning, or vehicle control.

Object tracking describes the task of estimating continuous trajectories of dynamic objects based on noisy and potentially fragmented sensor measurements. In the context of a classical automated vehicle system, tracking is situated within the perception stage of the processing chain [34]. The overall pipeline typically begins with sensor data acquisition (camera, LiDAR, radar, and Inertial Measurement Unit (IMU)), followed by detection and tracking of relevant objects. These perception outputs are fused into a coherent scene representation and forwarded to higher-level modules responsible for predicting future motion trajectories and planning safe paths for the ego vehicle. Finally, control modules translate the planning decisions into physical driving actions. Within this framework, object tracking is a critical enabler for downstream prediction and decision-making, as it ensures temporal continuity and contextual awareness of other traffic participants. In addition to linking detections over time, tracking also helps mitigate detection errors, such as false negatives or noisy localization, by integrating temporal context and state filtering. Furthermore, tracking enables the estimation of object-level properties that are not directly measured by visual sensors, such as an object's velocity or acceleration, which can be inferred through temporal integration.

Both classical and learning-based tracking approaches are considered in this work. The methodological scope includes classical formulations, which decompose tracking into detection, association, and state estimation, as well as learning-based, including end-to-end models.

This thesis focuses on visual tracking using cameras and therefore primarily adopts image-space supervision in the form of tight 2D bounding boxes in the image domain (see Figure 1.4). While LiDAR-based annotations in the form of 3D cuboids are available in some datasets and are used in selected experiments, projected 3D annotations are not assumed as the default representation for image-based tracking. Instead, 2D bounding boxes are used as they directly align with the image domain and avoid additional sources of uncertainty arising from sensor calibration, temporal synchronization, and geometric projection. Information from 3D, such as object positions or scene geometry, is considered complementary, for example to enable data generation pipelines, or to study multi-modal representations.

Within this scope and set of assumptions, the next section discusses the key challenges that arise.

1.1.4. KEY CHALLENGES FOR VISUAL TRACKING OF VRUs

Visual tracking of VRUs such as pedestrians and riders in urban traffic is affected by a range of external and internal factors. This section groups challenges into four broad categories: *vision-based limitations*, *challenges specific to object tracking*, *data annotation effort and scalability*, and *sensor fusion and system integration*.

Vision-based Limitations. Perception in real-world urban scenes is significantly impacted by both sensor-inherent limitations and challenging environmental conditions.

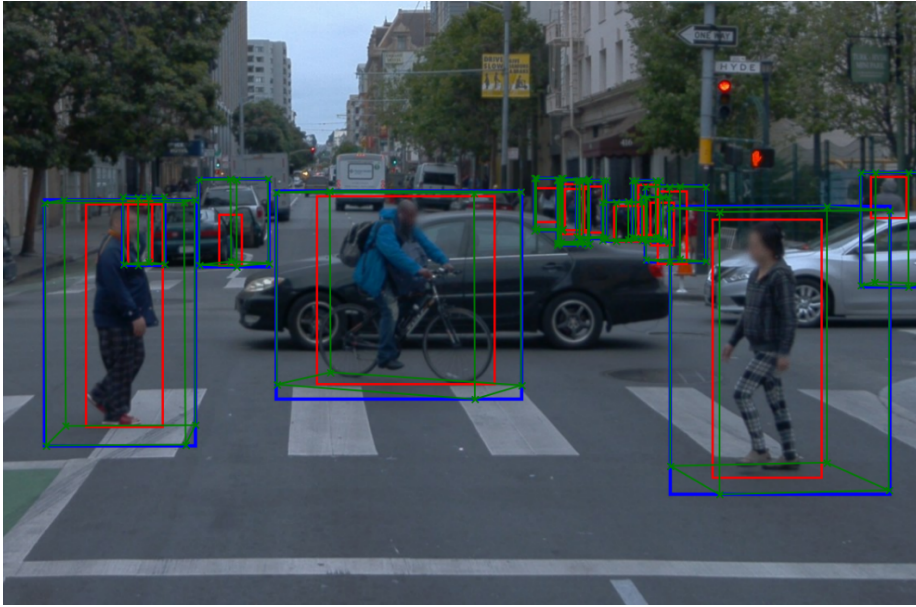


Figure 1.4: An example image from the WOD dataset [164]. The original 2D annotations are depicted in red, the projected 3D cuboid annotations as projected in the image (green) and the derived 2D bounding box (blue) for all persons larger than 40 pixels. As can be seen, the projection of 3D cuboids into 2D bounding boxes yield imprecise 2D annotations.

Sensor limitations include low spatial resolution, especially for distant or small targets (see Figure 1.5a). Pedestrians and riders captured at a distance often occupy only a few pixels, reducing the quality of appearance-based recognition and leading to degraded detection and identity association. Nighttime or poorly lit conditions further degrade the signal quality of image sensors, introducing noise, motion blur, and low contrast. This can severely affect both detection confidence and feature extraction.

Adverse weather such as rain, fog, or glare also impairs visual sensing. Cameras may suffer from water droplets or reflections on the lens, while fog and strong backlighting obscure scene content. See Figure 1.5b for examples. These external factors reduce visibility, increase false detections, and can lead to temporary loss of tracked targets.

In addition to environmental and sensor-based factors, the visual appearance of VRUs itself presents challenges. Pedestrians and riders can vary significantly in clothing, posture, and carried items (see Figure 1.5c). At the same time, visual similarity among individuals - such as similar clothing styles - makes identity association across time more difficult, particularly in crowded or homogeneous scenes. Furthermore, inter-class distinctions can be ambiguous in certain cases (e.g., a cyclist dismounting and walking), and human-like depictions such as mannequins or posters can introduce false positives, further increasing classification uncertainty.

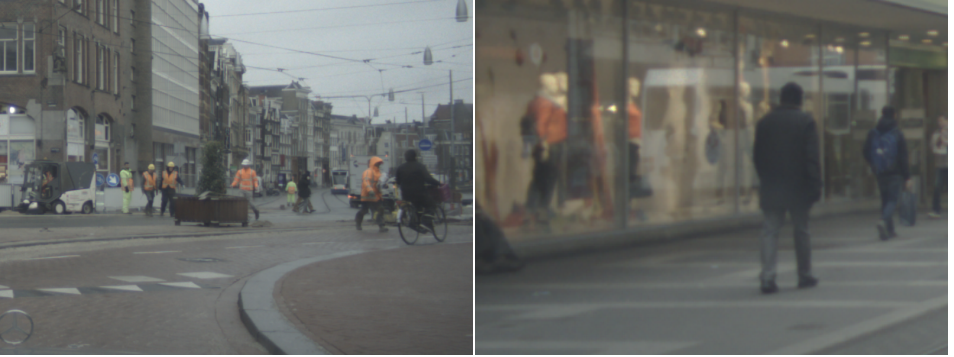
Challenges Specific to Object Tracking. Beyond detection, tracking VRUs across time introduces its own set of technical difficulties. Maintaining consistent identity over



(a) Examples showing sensor-inherent limitations. **Left:** Very distant pedestrians occupying only a few pixels in the image. **Right:** Scene during a turning maneuver at night, showing motion blur and low contrast.



(b) Example of adverse weather and lighting conditions. **Left:** Rainy scene with visible water droplets on the windshield and a wiper in motion. **Right:** Sunny scene with strong sun glare obscuring parts of the view.



(c) Examples of challenging [VRU](#) appearance variations. **Left:** Road workers with highly similar appearance (yellow helmets, dark trousers, warning vests). **Right:** Mannequins in a shop window depicting real pedestrians.

Figure 1.5: Challenging visual conditions for vision-based [VRU](#) perception in urban driving. All examples are from the EuroCity Persons data recordings.



(a) A crowded nighttime pedestrian zone with many VRUs and partial occlusions, creating frequent changes in visibility and association difficulty.



(b) A man changes from kneeling to standing within 1.5 seconds, altering his appearance and bounding box geometry



(c) A woman emerges between parked cars, briefly stops, and then runs across the street, challenging motion models.

Figure 1.6: Tracking-specific challenges for VRUs. In each example, the two images are 1.5 seconds apart. All examples are from the EuroCity Persons data recordings.

time is crucial for downstream tasks such as prediction and risk estimation. However, in crowded scenes (see example in Figure 1.6a) or at complex intersections, VRUs frequently occlude each other or become temporarily hidden behind other objects such as parked vehicles or roadside infrastructure. This leads to identity switches, fragmented trajectories, or track losses.

Appearance and pose variability pose another critical challenge. VRUs differ in clothing, body size, posture, and movement patterns. These variations are amplified by dynamic changes during walking, cycling, or object interaction. Such intra-class variability complicates feature extraction and identity matching. Additionally, the non-rigid nature of human motion results in frequent shape and pose deformations, adding variability to both detection bounding boxes and extracted features, as can be seen in Figure 1.6b.

Low detection confidence or misclassifications can propagate through tracking pipelines, especially in systems that rely heavily on frame-level detections. Without robust temporal integration and ReID mechanisms, tracking systems are prone to drift or failure in complex scenes. High-quality tracking therefore demands models that can bridge detection gaps, recover from partial occlusions, and reason over longer temporal windows.

In dynamic traffic environments, VRUs exhibit sudden and unpredictable changes in motion, such as abruptly changing walking direction, accelerating on a bicycle, or reacting to a nearby vehicle (see Figure 1.6c). These behavioral patterns reduce the effectiveness of simple motion models and challenge state estimation modules that rely on temporal smoothness.

Data Annotation Effort and Scalability. Aside from these technical challenges, object tracking places substantially higher demands on annotated data than related tasks such as object detection. Tracking datasets not only require accurate spatial localization, but also reliable temporal correspondences across timestamps, typically at higher frequencies and over longer continuous sequences. In urban traffic scenes, where frequent occlusions, interactions, and re-identification events occur, this significantly increases the effort and complexity of manual annotations. As a result, the creation of large-scale tracking datasets is considerably more expensive and time-consuming than the annotation of frame-wise detection datasets, limiting the availability of training and evaluation data.

Sensor Fusion and System Integration. Modern perception systems increasingly adopt multi-modal sensor setups, combining camera and LiDAR to compensate for the limitations of each. While this fusion improves robustness, it introduces new integration challenges. LiDAR data, while less sensitive to lighting, becomes sparse at greater distances and provides limited semantic information. Cameras, on the other hand, offer rich texture but struggle in darkness. Combining these modalities requires precise time synchronization and geometric calibration, in particular accurate extrinsic calibration that defines the relative sensor placement between camera and LiDAR. Calibration drift refers to a change in this relative sensor placement over time. Together with sensor noise or projection errors, such drift can lead to inconsistencies that degrade fusion performance and introduce ambiguity in tracking or classification results.

1.1.5. RESEARCH QUESTIONS OF THIS THESIS

The previous section has highlighted a set of challenges that jointly shape the problem of visual tracking of VRUs in urban automated driving. Building on this analysis, the aim of this thesis is not to resolve each challenge in isolation, but to investigate how tracking performance can be improved under the combined constraints identified in the previous section. Accordingly, the thesis is guided by the overarching Research Question (RQ):

Main RQ: *How can visual tracking of vulnerable road users in urban traffic environments be improved under data, annotation and sensor constraints?*

To address this main question in a structured manner, the thesis formulates a finer-grained set of research questions. The resulting research questions are summarized as follows:

- **RQ1:** What methods are used for visual tracking and person re-identification of VRUs, and what limitations emerge when these methods are considered in the context of urban automated driving?
- **RQ2:** How can temporally dense trajectory supervision be generated from sparse 2D annotations, using only on-board sensors?
- **RQ3:** How reliable is the generated pseudo ground-truth for training and benchmarking modern tracking methods?
- **RQ4:** Can person re-identification be improved by combining information from camera and LiDAR?

These sub-questions reflect how the challenges in Section 1.1.4 manifest at different levels of the tracking problem. Specifically, the thesis first examines the capabilities and limitations of existing visual tracking and person re-identification approaches in the context of ego-centric urban driving scenarios (RQ1). Following this, the thesis turns to challenges related to data availability and annotation effort. It investigates how the substantial data demands of tracking can be met with reduced manual annotation effort (RQ2), while explicitly relying on pseudo ground-truth generated in a semi-supervised manner. Addressing this challenge naturally raises the question of whether the resulting pseudo ground-truth is sufficiently reliable for training and benchmarking tracking methods (RQ3). Finally, beyond data-related constraints, the challenges discussed in Section 1.1.4 also point to limitations of purely vision-based perception. In this context, the thesis explores whether complementary sensing modalities can be leveraged to strengthen identity association (RQ4).

Section 1.2 provides an overview of the individual chapters and explains how each chapter addresses one or more of the above research questions.

1.2. THESIS OUTLINE AND CONTRIBUTIONS

This thesis aims to improve the safety of VRUs by advancing robust multi-object tracking and person ReID for automated driving systems. The primary focus lies on camera-based perception, with LiDAR data used in a complementary manner where beneficial.

The research targets challenging urban environments, where VRUs are particularly at risk. The thesis is organized into five chapters to address the challenge of visual tracking and identification of VRUs in urban automated driving settings.

This section outlines the structure of the thesis and explains how the individual chapters contribute to answering the RQs introduced in Section 1.1.5. Chapter 2 addresses RQ1 by reviewing the previous work in object tracking and person re-identification, covering both traditional model-based methods and recent developments in deep learning. It further discusses strategies for dataset generation and surveys existing tracking benchmarks in automated driving. This sets the stage for the data-driven approaches pursued throughout this thesis and highlights the critical dependency of modern methods on large-scale annotated datasets. Recognizing the shortage of such datasets, Chapter 3 addresses RQ2 by presenting a novel semi-supervised pipeline for generating dense person trajectories from sparsely labeled images. By leveraging detection outputs, geometric constraints, and batch-level processing, this pipeline enables scalable annotation of training data without requiring exhaustive manual effort. Building on this methodology, Chapter 4 introduces the EuroCity Persons 2.0 (ECP2.0) dataset - a large-scale benchmark for urban pedestrian and cyclist tracking generated using the pipeline presented in Chapter 3. The dataset incorporates sequences from diverse environments. Chapter 4 addresses RQ3 by evaluating the reliability of the generated pseudo Ground-Truth (GT) for training and benchmarking. Chapter 5 addresses RQ4 by investigating whether person re-identification performance can be improved through the integration of camera and LiDAR information. Finally, Chapter 6 reviews the findings of the thesis by revisiting the RQs and by outlining limitations and directions for future work.

The following subsections provide a detailed overview of each chapter and highlight the specific contributions of this thesis in more detail.

1.2.1. PREVIOUS WORK: TRACKING WITH DEEP LEARNING

The aim of Chapter 2 is to give the reader a brief yet comprehensive introduction to the field of object tracking with a focus on the latest deep-learning based extensions and approaches. At first, traditional non-deep tracking systems are briefly reviewed and a generic model of the individual components of such systems is introduced. Based on this structure the representative deep-based tracking applications in the literature are classified and presented, including the shift to end-to-end formulations and transformer-based architectures. Beyond algorithmic paradigms, the chapter reviews approaches for dataset generation and surveys existing benchmarks for tracking in automated driving. Finally, it introduces related work on person re-identification as a complementary task to multi-object tracking.

Section 2.1 and 2.2.1 are based on the work published in:

- [75] S. Krebs, B. Duraisamy, and F. Flohr “A survey on leveraging deep neural networks for object tracking”. In Proc. of the IEEE Intelligent Transportation Systems Conf. Workshops (ITSC), 2017, pp. 411–418 (©2017 IEEE)

Contribution of this chapter:

- Provides a survey on tracking with deep learning by summarizing the progression toward deep-learning-based tracking models, including early modular ap-

proaches and the rise of end-to-end formulations. The chapter establishes the conceptual groundwork for the thesis and motivates the need for data-driven methods addressed in later chapters.

1.2.2. SEMI-SUPERVISED CREATION OF AN AD PERSON DATASET

Reliable deep learning-based multi-object tracking in automated driving requires large amounts of temporally consistent training data. However, producing such annotations manually is costly and labor-intensive. This chapter addresses this challenge by exploring how sparse human annotations can be expanded into dense training trajectories in a semi-supervised manner.

The presented pipeline is designed to leverage parts of the on-board sensor setup typically available in automated vehicles - namely, cameras, LiDAR, and inertial sensors. Starting from sparse 2D bounding box annotations at selected keyframes, image-based bounding boxes and projected LiDAR point clouds are used to estimate the 2.5D location of each object through a constrained clustering and label uplifting process.

To produce temporally complete trajectories, two alternative batch-wise approaches are presented. The first is a graph-based optimization strategy that links detections across time while accounting for noise, occlusion, and false positives. The second is a modular two-stage tracklet-based method. It operates on tracklets, which are short and reliable local track segments, and merges them based on spatial and temporal coherence. Both approaches incorporate ego-motion compensation to ensure consistent spatial alignment and apply model-based smoothing for trajectory refinement.

An experimental evaluation of the proposed pipeline compares the two trajectory generation approaches against baseline methods, complemented by ablation studies and a ceiling analysis.

Chapter 3 is based on the works published in:

- [74] S. Krebs, M. Braun, and D. M. Gavrila “*Generating 3D Person Trajectories from Sparse Image Annotations in an Intelligent Vehicles Setting*”. In Proc. of the IEEE Intelligent Transportation Systems Conf. (ITSC), 2019, pp. 783–788 (©2019 IEEE)
- [11] M. Braun, S. Krebs, and D. M. Gavrila “*ECP2.5D - Person Localization in Traffic Scenes*”. In Proc. of the IEEE Intelligent Vehicles Symposium (IV), 2020, pp. 1694–1701 (©2020 IEEE)
- [73] S. Krebs, M. Braun, and D. M. Gavrila “*EuroCity Persons 2.0: A Large and Diverse Dataset of Persons in Traffic*”. In IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), 2024, vol. 46, no. 12, pp. 10929–10943 (©2024 IEEE)

This chapter's main contributions are:

- Introduction of a semi-supervised annotation pipeline that generates dense person trajectories from sparse 2D image annotations, leveraging typical automated vehicle sensor setups (camera, LiDAR, inertial sensors).
- A method for label uplifting that estimates 2.5D person locations through constrained clustering of projected LiDAR points aligned with 2D image measurements.

- Two alternative trajectory generation approaches: First, a batch-wise graph optimization approach for global association and second, a modular tracklet-based pipeline for efficient temporal linking and smoothing.

1.2.3. ECP2.0 DATASET FOR PERSON TRACKING IN AD

Chapter 4 presents the EuroCity Persons (ECP) 2.0 dataset, a novel image dataset for person detection, tracking and prediction in traffic. The dataset was collected on-board a vehicle driving through 29 cities in 11 European countries. It contains more than 220K unique person trajectories, in more than 2.0M images. ECP2.0 is about one order of magnitude larger than previous state-of-the-art person datasets in automotive context. It offers remarkable diversity in terms of geographical coverage, time of day, weather and seasons.

The chapter details the dataset's structure, annotation characteristics, and coverage, and provides a statistical overview of object distribution, track lengths, and spatial properties. In addition to its descriptive components, the chapter presents a series of experiments validating the quality and utility of the generated pseudo ground-truth, with benchmarking experiments showing how the dataset improves the performance of state-of-the-art tracking models. The ECP2.0 dataset is made publicly available for non-commercial research purposes.

This Chapter is based on the work published in:

- [73] S. Krebs, M. Braun, and D. M. Gavrilu "EuroCity Persons 2.0: A Large and Diverse Dataset of Persons in Traffic". In IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), 2024, vol. 46, no. 12, pp. 10929–10943 (©2024 IEEE)

The contributions of this chapter are twofold:

- Introduces ECP2.0 and outlines its key characteristics, including its large scale, rich diversity, and the use of semi-supervised annotations derived from the pipeline in Chapter 3.
- Validation through benchmark experiments demonstrating the dataset's quality and suitability for training and evaluation of modern deep learning-based tracking methods.

1.2.4. LEARNING APPEARANCE EMBEDDINGS FOR PERSON REID

Chapter 5 introduces a novel method for creating appearance embeddings to identify individual persons using an object re-identification framework. CLFormer (Camera LiDAR Transformer) is presented, a transformer-based architecture that incorporates multi-modal data from both camera and LiDAR sensors. The 3D Cuboid-Inclusive Point Embedding (3D-CIPE) is introduced, which leverages rich data from LiDAR point clouds and 3D cuboids to add a learnable embedding into the transformer structure. Additionally, through ablation studies, various strategies for the early and late fusion of multi-modal input data are explored and analyzed.

To evaluate the proposed CLFormer, the nuScenes dataset [14] is reinterpreted for ReID purposes and used in the experiments. The method demonstrates a significant improvement in performance, outperforming the image-only baseline with an increase

of 2.3 in mean Average Precision (mAP).

This Chapter is based on the work published in:

- [76] S. Krebs and D. M. Gavrilă “*Camera- and LiDAR-based Person Re-identification*”. In Proc. of the IEEE Intelligent Vehicles Symposium (IV), 2025, pp. 1408–1414 (©2025 IEEE)

The contributions of this chapter are threefold:

- Introduces CLFormer, a transformer-based ReID architecture that combines camera and LiDAR data.
- Proposes the 3D Cuboid-Inclusive Point Embedding (3D-CIPE) for incorporating LiDAR and 3D cuboid information.
- Investigation and comparison of different sensor fusion strategies for effectively combining LiDAR and image data within the ReID framework.

2

PREVIOUS WORK

This chapter reviews previous work on vision-based multi-object tracking and person re-identification, with a particular focus on vision-based approaches relevant to intelligent vehicles. It begins with traditional modular tracking frameworks and then surveys recent deep learning-based methods. In addition, strategies for dataset generation (Chapter 3), tracking datasets in the automated driving domain (Chapter 4), and related work on person re-identification (Chapter 5) are presented.

2.1. TRADITIONAL OBJECT TRACKING

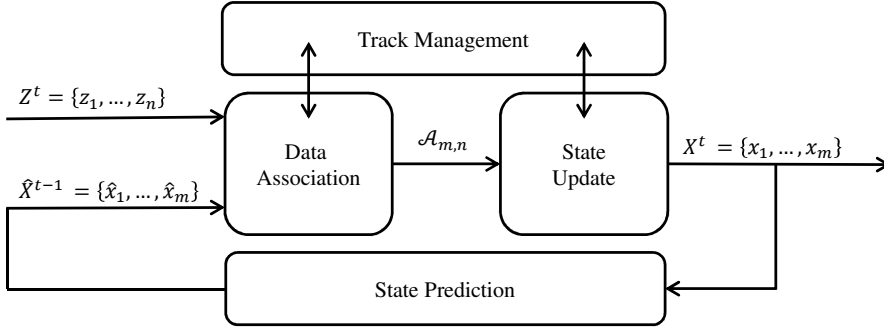


Figure 2.1: Schematic illustration of the single components of a generic traditional multi-object tracking framework. The set of n new measurements Z^t at each time t is used to update the predicted states $\hat{X}^{t-1}$ of the m already tracked objects, based on the calculated association matrix $\mathcal{A}_{m,n}$.

Early development in object tracking emerged primarily from military applications such as radar or sonar tracking [4]. In general, the aim of these algorithms is to estimate the current state x_i^t at time t of the i -th object, by using a set of noisy measurements $Z^t = \{z_1, \dots, z_n\}$. These tracking methods can be roughly separated into the four coherent

parts of: Data Association, State Update, State Prediction, and Track Management, as illustrated in Figure 2.1. Although these components can be conceptually separated, they are often tightly coupled in practical implementations.

Data Association. This component describes the problem of finding the correct correspondences between the set of already known tracked targets and the ensemble of new possible noisy measurements. To this end, a distance or cost function is typically defined to quantify the likelihood that a given measurement matches a particular track. Early methods such as the Multi Hypothesis Tracking (MHT) [8] and the Joint Integrated Probabilistic Data Association (JIPDA) [4] integrate this association step tightly within their frameworks: JIPDA computes weighted updates based on association likelihoods, while MHT maintains multiple possible association hypotheses over time in a tree structure. A more general approach involved constructing an association matrix using costs or distances between all track and measurement pairs, and solving the resulting assignment problem using greedy matching or Hungarian algorithm [77]. Especially for visual tracking the use of an additional appearance model, which encapsulates the visual information of the tracked object has yielded significant performance increases during data association.

State Update. Once associations between measurements and tracks have been established, the estimated states of the objects are updated accordingly. During the update of the object's state, the uncertainty of the current state as well as the potential sensor noise have to be taken into account. One of the most utilized and popular methods to perform this estimation recursively is the Kalman Filter (KF) [69]. The Kalman Filter provides a closed-form Bayesian estimation solution under linear-Gaussian assumptions and is optimal for linear systems with Gaussian noise. In the presence of non-linear problems, several extensions of this approach have been presented, namely the Extended Kalman Filter (EKF) and the Unscented Kalman Filter (UKF). For non-linear and/or non-Gaussian systems, Particle Filters (PFs) are a widely used choice. PF refers to a family of sequential Monte Carlo methods that approximate a numerical solution to the Bayes recursion, representing the posterior distribution with a set of weighted particles. Besides using PFs solely to estimate the state distribution of the object, plenty of published work incorporates data association into the filtering process [31, 145]. These Random Finite Set (RFS)-based tracking techniques model the multi-object state as a (potentially) multi-modal random finite set and often employ particle-based approximations. In such formulations, data association is handled implicitly via the multi-object likelihood within the update step, avoiding explicit one-to-one assignments.

State Prediction. Closely related and usually entangled with the state update is the state prediction. Most trackers operate in discretized time cycles. Since the world is not static and typically either the object position, the sensor position, or both change over time, these changes must be accounted for between tracking cycles. Consequently, the target's state is propagated from the time of the last update to the time at which new measurements become available. These predictions are primarily derived from the object's kinematic state and may incorporate additional a-priori knowledge about the

expected motion patterns [88]. The specific choice of motion model depends strongly on the application context and the type of object being tracked. While simple models such as constant-velocity or constant-acceleration are sufficient in many cases, more advanced models may be required to capture specific behaviors like turning or stopping. Generally, the further the prediction extends into the future, the greater the associated uncertainty becomes.

Track Management. This component governs the life cycle of object tracks, including track initialization, confirmation, and deletion. Tracks are typically created when a sequence of unassigned measurements exhibit spatial and temporal coherence, suggesting the presence of a previously untracked object. However, a newly created track is often considered tentative until it receives consistent support over multiple time steps. To manage false positives and ensure track stability, many systems adapt heuristic mechanisms such as finite-state machines, which transition tracks between states like tentative, confirmed, and deleted. Tracks are deleted when no associated detections are received over a predefined period, indicating that the object is no longer observable or has left the scene.

2.2. TRACKING WITH DEEP LEARNING

This section provides a structured overview of the use of deep learning for object tracking, with a particular focus on its relevance for automated driving applications. A distinction is made between methods developed until 2017 (Section 2.2.1) and those published after 2017 (Section 2.2.2).

Section 2.2.1 is based on the survey paper [75] from 2017, which summarizes the initial wave of deep learning-based tracking research. These early contributions typically extend classical modular tracking pipelines with deep learning components, for instance by replacing hand-crafted features with CNN-based embeddings. Thus, this section follows the traditional structure introduced in Section 2.1 as many methods keep the classical components separate and classifies them into: **feature**, **data association**, **prediction**, and **end-to-end** learning.

In contrast, Section 2.2.2 covers methods published since 2017, reflecting a paradigm shift toward fully trainable and end-to-end systems. These later approaches frequently unify detection, embedding, or tracking within single architectures, leverage attention mechanisms (e.g., transformers), or employ joint optimization strategies. Rather than decomposing the pipeline into independent modules, modern methods tend to be categorized by how the tracking problem is formulated. Accordingly, Section 2.2.2 groups methods into three dominant paradigms: **Tracking-by-Detection**, **Tracking-by-Regression**, and **Tracking-by-Attention**. These categories highlight the increasing integration of detection, association, embeddings, and motion reasoning within unified frameworks.

2.2.1. DEEP LEARNING FOR TRACKING: UNTIL 2017

Emerging from the field of visual tracking, the first deep-learning-based approaches leveraged the robust feature representations learned by the networks to better model the

tracked objects. These approaches [27, 105, 178–180] often use deep neural networks as a black-box feature extractor and perform the subsequent classification (same-object, other-object), association and filtering with more traditional methods. To overcome the challenges caused by the non-deep successive classification and association parts, some approaches [84, 167, 170] aim to solve the correspondences problem in a deep learning manner. These methods mostly employ Siamese networks [28] to directly learn a similarity measure given two input patches containing the possible object match candidates. Based on the calculated match probabilities, well-known methods such as the Hungarian algorithm [77] can calculate the optimal correspondences.

Further leveraging the power of deep neural networks to capture non-linear dependencies directly from the data, some methods [2, 60, 199] focus on learning to predict the movement of objects. The knowledge of the networks which was gathered during the offline training is used to estimate the most likely future positions of all objects given their history.

Lastly, by formulating tracking as an end-to-end problem some approaches introduce networks which jointly optimize the whole tracking process. These works [46, 58, 114, 118, 121] unite all above-mentioned traditional components of a traditional tracking pipeline into one network.

USING DEEP LEARNING FOR FEATURE LEARNING

Initial research efforts focused on utilizing the powerful and robust image features learned by deep networks to model the tracked object effectively.

One of the earliest approaches is the Deep Learning Tracker (DLT), introduced by Wang and Yeung [180] in 2013. In their work the authors train a stacked Denoising Autoencoder (DAE) offline on a large dataset of auxiliary natural images to obtain a generic object feature representation which is transferred to the online tracking process. The use of a DAE increases the robustness of the learned image features, compared to a standard Auto Encoder (AE). For the online tracking process, the encoder part of the stacked DAE is used as a pre-trained feature extractor extended by a sigmoid classification layer. The DLT uses a particle filter as the motion model and to perform the state estimation. During tracking for each frame, a set of particles is drawn from the currently estimated patch. The confidence of each particle is calculated by making a simple forward pass through the network. If the sum of the calculated confidences for a frame falls below a pre-defined threshold, the network needs to be re-tuned again to cope with the changing appearance of the tracked object. However, the employed network structure limits the input image size, causing blurry features, and the linear classification layer does not yield optimal results. To overcome these issues, Wang et al. presented the Structured Outcome Deep Learning Tracker (SO-DLT) [179] which is an enhancement of the DLT. Instead of the stacked DAE a novel structured output CNN is used during pre-training to distinguish between objects and background. The CNN outputs a pixel-wise probability map indicating whether the pixel belongs to the bounding box of an object. Especially due to the classification performance of the CNN, the results of the SO-DLT are superior to those of the DAE-based DLT approach.

In their work, Ma et al. [105] also use a CNN which is trained offline on the large-scale ImageNet dataset with category-level labels. The authors state in their paper that higher level convolutional layer in the CNN encodes the semantic object information. However,

their spatial resolution is too coarse to precisely infer the target location. The feature maps from lower level layers provide more precise position information, but lack the semantic information needed for appearance-invariance. Thus, reasoning with multiple layers of different levels of CNN features is essential for target tracking. In [105], three different convolutional layers from the pre-trained VGG-net [157] are used. The outputs of these layers are implemented as multi-channel features to learn an adaptive linear correlation filter per layer. During the online tracking the generated feature maps of each layer are convolved with the respective learned correlation filter to obtain a response map. Using a coarse-to-fine search approach the object location is inferred based on the calculated multi-level response maps.

Similarly, Wang et al. [178] conducted an in-depth feature analysis on their ImageNet pre-trained VGG-based CNN with the same findings. Further, the authors state that only a subset of the neurons is relevant for tracking a specific target. Therefore, a tracking framework is introduced with a feature map selection method that avoids irrelevant feature maps as well as over-fitting on noisy ones. The selected feature maps are jointly considered at two different levels, by two separate networks (the general network GNet and the specific network SNet). Both networks perform a heat map regression to infer the object location.

In the work of Hong et al. [61] a pre-trained CNN is used for target localization and feature extraction. The extracted features are classified by an online learned Support Vector Machine (SVM). The features of each positive classified sample are back-propagated through the network to obtain a saliency map. By combining the saliency maps produced by each sample, an object-specific spatial saliency map can be calculated. Tracking is performed by sequential Bayesian filtering using the object-specific map as observation along with an online learned appearance model, resulting in the final object probability map. Based on the calculated track, new samples are extracted which are employed for the online-training of the SVM.

USING DEEP LEARNING FOR DATA ASSOCIATION

Data association remains a core challenge in multi-object tracking. Deep networks have been adopted to learn similarity metrics directly from data, eliminating the need for hand-crafted cost or distance functions. Siamese networks are a common choice for this purpose. These twin-networks consist of two identical sub-networks joined at the output.

Tao et al. introduce the Siamese INstance Search Tracker (SINT) [167]. The main concept of the SINT is the use of a two-streamed Siamese deep neural network, which is explicitly designed for tracking. During the training phase, the network is provided with two data streams: the search and the query stream. The query stream provides a video frame with the exact position of the tracked object, whereas another - not necessarily adjacent - frame is given by the search stream with randomly sampled object locations. The random object proposals are used as positive training samples if they exceed a certain overlap-threshold with the desired object, otherwise they are used as negative samples. Thus, the Siamese network learns a generic measure, which for a given pair of input images expresses the probability that the patches contain the same object. During tracking, the position of the object in the first frame is given. This initial position is used to extract

a template of the tracked object. During the entire tracking sequence, this template is utilized to calculate the similarity of the new candidate regions. The candidate regions are created using a radius sampling approach. Finally, the region with the highest similarity is considered the new object position and the box positions are further refined by four Ridge regressors.

Leal-Taixe et al. [81] also leverage the ability of a twin-architecture Siamese network to learn a similarity measure directly from data. In their work, the authors focus on pedestrian tracking. The network is trained to learn local spatio-temporal features to distinguish people, by aggregating pixel values with optical flow. Further, the output of the learned network is extended by a set of contextual features - which capture the relative position and size change of the objects, as well as their relative velocity - using gradient boosting. The learned likelihood function is used by a modified linear programming approach to obtain the final multi-object pedestrian tracker.

Instead of using the optical flow to model the spatio-temporal dependencies, Varior et al. [170] use a recurrent version of the Siamese network. The introduced LSTM-based Siamese network can memorize spatial dependencies over time, and is further able to selectively propagate relevant information. This enhances the discriminate capabilities of the network in the context of person re-identification, and optimizes the employed contractive loss function (inter-pair distance increases, whereas intra-pair distance declines).

USING DEEP LEARNING FOR PREDICTION

Traditional tracking systems rely on hand-crafted motion models. Deep networks offer an alternative by learning such models directly from data.

To this extent, Li et al. propose the Behavior-CNN [199] to model the kinematic properties of pedestrians in crowded scenes. The employed CNN is trained with real-world data from a surveillance scenario involving a static camera. The authors perform an in-depth investigation on different aspects of the learned Behavior-CNN. They state that the learned location maps represent the location awareness of the network for the learned scene (note that this is only possible due to the static sensor pose). Furthermore, the feature maps of the trained layers are examined, resulting in the low-level layers capturing simple motion patterns such left or up movements, while higher-level layers capture more semantic object movements which also influence other objects. Lastly, the impact of the receptive field size is reviewed, which is crucial to capture inter-object dependencies.

Alahi et al. also focus on the prediction of human trajectories in multi-person scenarios. Therefore, they introduce the Social-LSTM [2] which interprets the task of predicting the future trajectories of multiple persons as a sequence generation problem. For each person present in the scene a single LSTM is utilized to predict the future trajectory based on the pre-processed trajectory history. To also be able to capture inter-class dependencies, a novel social pooling layer is introduced, which enables the Social-LSTM to capture connections between multiple spatially proximal persons. This renders the LSTM-based approach feasible to jointly predict multiple human trajectories precisely.

Lastly, Hoermann et al. [60] focus in their work on the prediction of a Dynamic Occupancy Grid Map (DOGMa) for an automated vehicle in an urban driving scenario using

deep learning. The DOGMa is used as an intermediate time-filtered environment representation obtained from the raw sensor measurements from multiple sensors. The map is a probabilistic grid-based representation of the surrounding static and dynamic objects, with probability of allocation for each cell. Using an image-like representation of the map as input for the employed neural network allows predicting the perceivable scene. During training, a fully automatic label generation is used to overcome the need for annotating the recorded real-world data manually. It is shown that the trained network can correctly predict complex scenarios at an urban intersection involving multiple road users (pedestrians, cars, cyclists) with different maneuver classes.

END-TO-END LEARNING FOR TRACKING

The superior results of most deep-learning based visual detection methods can be attributed to the fact, that instead of using hand-crafted features, the features are learned directly from the data. By using an end-to-end approach this could also enable a performance boost for the task of object tracking, since the full tracking pipeline - object representation, object extraction and location prediction - is jointly tuned to maximize the overall tracking quality.

By using a recurrent neural network Gan et al. [46] present an anonymous model-free visual object tracking approach. The network is completely trained offline with synthesized data which simulates a moving object with changing velocity, acceleration and direction. The recurrent neural network is used to capture the spatio-temporal dependencies and fuse past predictions with their corresponding visual features. During the training phase the artificial raw video frames are used as input, along with the ground-truth location of the desired object in the first frames. This implicitly enforces a model-free anonymous tracking, since no further knowledge of the object is given. The network calculates the estimated bounding box for the tracked target for each frame as its output.

A similar approach is presented by Held et al. [58] with the Generic Object Tracking Using Recurrent Networks (GOTURN). The network is trained entirely offline to learn a generic relationship between motion and appearance of novel objects. As stated by the authors, the network is capable of tracking an unknown object with up to 100 Frames per Second (fps) demonstrating high efficiency in single-object tracking.

Another end-to-end based approach is presented by Nam and Han [117], called the Multi-Domain Network (MDNet). The essence of the work is to separate domain-specific knowledge from domain-independent one. Therefore, the MDNet is composed of one shared layer branch which is utilized to learn a generic feature representation employable to tracking, and several domain-specific branches. Each domain-specific binary classification branch is trained only with a set of videos containing objects of the same class. To obtain the final tracking network, all binary domain branches are removed and replaced by a new classification layer, which is trained online during tracking to adapt to the desired domain.

Ning et al. [118] extend the object detection network You Only Look Once (YOLO) [143] in a recurrent fashion, creating the Recurrent YOLO (ROLO). Their proposed network concatenates the high-level features produced by the general object detection network with a Long-Short Term Memory (LSTM) to model the temporal relation along with the spatial information. Thus ROLO directly allows end-to-end spatio-temporal regression of either object coordinates or heat maps. The authors state that the network is

"double deep" (temporal and spatial) since it examines the history of all previous object locations as well as the former robust visual features. The detection part of the network is pre-trained on the ImageNet dataset, whereas the training of the tracking module is performed in a second stage with the ADAM stochastic optimization method.

The online multi-target tracking framework presented by Milan et al. [114] is closely related to traditional target tracking and can solve the data association, state estimation and prediction, as well as the creation and deletion of tracks. The network consists of a traditional recurrent neural network part which is employed for state update, state prediction, and existence probability estimation. The state space is modeled as multi-dimensional with continuous and discrete variables. While the recurrent neural network performs well for the above mentioned tasks it cannot properly handle the combinatorial problem of data association. Thus, an LSTM-based architecture capable of solving the highly complex one-to-one assignment problem entirely from offline training is designed.

Contrary to most approaches Posner and Ondruska [121] assume a complete yet uninterpretable state of the world for their Deep Tracking framework. By leveraging RNNs the authors are able to train the network end-to-end with partial observations of dynamic scenes. The offline training is performed using a novel dropout method, in which observations are dropped spatially across the entire scene, as well as temporally across multiple time steps. Based on the raw sensor measurements obtained from a 2D laser-sensor, the network is able to predict the non-occluded state of the entire scene as an occupancy grid map based representation. The approach has proven applicable to synthetic data. The Deep Tracking architecture was extended by Ondruska et al. [120]. To increase the performance especially for challenging real world scenarios, multi-scale convolutions are used to cope with objects of different sizes. Further static memory is employed to learn spatial information, and dynamic memory is applied so that information can be used for a longer period of time. Finally, due to inductive transfer, the tracking features learned during training are used to extend the network to be able to classify the tracked objects. The resulting recurrent neural network, tailored to tracking in complex and dynamic scenarios, is empirically evaluated with real-world data collected at a busy road intersection. Finally, the Deep Tracking framework was further extended by Dequaire et al. [36] to be applicable to tracking objects from an automated vehicle. Since the vehicle itself moves in the world, the ego-motion needs to be corrected and decoupled from the objects' motion. This is achieved by introducing a Spatial Transformer Module (STM) in the hidden state of the network, which performs a transformation on all feature maps based on an external received ego-motion.

2.2.2. DEEP LEARNING FOR TRACKING: SINCE 2017

This section reviews related work on multi-object tracking, focusing on developments since 2017. Earlier contributions to the field - prior to 2017 - are covered in Section 2.2.1. The goal of this section is to provide a structured overview of the field by organizing existing approaches into three main categories: tracking-by-detection, tracking-by-regression, and tracking-by-attention. These categories reflect the dominant paradigms in current research and differ in how they formulate the tracking task, whether through explicit detection and association, direct regression, or query-based attention mecha-

nisms. Relevant sub-categories and trends, such as appearance-based matching, joint detection and embedding (JDE), motion-centric refinements, and the fusion of LiDAR and camera data, are discussed within each section.

TRACKING-BY-DETECTION

Without Appearance Features. The tracking-by-detection paradigm has long been a dominant approach in Multi-Object Tracking (MOT), gaining increased popularity with the improvement of detection algorithms driven by deep learning and large-scale datasets. As the name suggests, tracking is performed by linking detections - generated independently per frame - into consistent trajectories over time. The core of these methods lies in data association, which is typically formulated as a matching problem between detections and existing tracks using a cost function. SORT [7] established this approach with a Kalman filter to update individual tracks and the Hungarian algorithm [77], using Intersection-over-Union (IoU) as a cost function. Extensions have focused on improving the association strategy, modifying cost functions, or refining the underlying motion model to better handle real-world challenges such as occlusions and abrupt motion changes.

OC-SORT [16] improves upon SORT by introducing two key contributions. First, Observation-Centric Re-Update (ORU) replaces Kalman filter-based prediction during occlusion gaps with linear interpolation between observations. Second, Observation-Centric Momentum (OCM) introduces a velocity-based cost to better align moving targets with similar directionality, especially helpful in dynamic scenes.

Focusing on the association strategy, BYTETrack [209] proposes a two-stage association strategy: high-confidence detections are first associated with active tracks, and remaining unmatched tracks are then considered for low-confidence detections. This simple idea led to substantial improvements in handling occlusion and low-score detection drift, rendering BYTETrack as a widely adopted baseline [1, 97, 106, 162, 176]. BYTEv2 [162] extends this idea further by incorporating suppressed detections from Non-Maximum Suppression (NMS) and improving camera motion compensation. Further association focused work includes SparseTrack [97], which proposes a pseudo-depth-based association mechanism to improve robustness in crowded scenes. Detections and tracks are clustered by estimated pseudo-depth and matched preferentially within the same cluster, reducing the likelihood of mismatches across overlapping targets. TrackMPNN [138] shifts the focus from hand-crafted association to a learned graph-based method. A dynamic graph encodes object relations, and message passing is used to infer association probabilities.

In the context of LiDAR-based or 3D MOT, Complexer-YOLO [156] extends LiDAR-based 3D detection to tracking by applying a random finite set (RFS) formulation. BYTETrackv2 [211] extends BYTETrack to 3D, MOLTR [86] combines monocular 3D object detection and shape reconstruction, while MSAMOT [222] uses multi-stage matching across 2D and 3D modalities. The PIoU tracker [189] proposes a pose-aware IoU metric using torso-only bounding boxes to better handle person deformation.

With Appearance Features. In many real-world tracking scenarios, motion or spatial cues alone are insufficient to maintain consistent identities over time. This is particu-

larly true in crowded scenes, during occlusions, or when objects follow complex motion patterns. To address these challenges, appearance information can be incorporated as an additional cue for association. By comparing visual similarity between detections across frames, appearance-based methods can help distinguish between targets that are spatially close or have been occluded. This additional modality can enhance robustness and reduce identity switches, especially in pedestrian tracking tasks where people often move in groups or cross paths.

Among the first methods, DeepSORT [187] extends the basic SORT by incorporating appearance embeddings for data association. The appearance embeddings are generated by a separate convolutional neural network and are used to compute similarity scores during association. A large number of follow-up works extend or modify this architecture.

Appearance-enhanced methods include BoT-SORT [1], which extends BYTETrack [209] by appearance embeddings while also introducing ego-motion compensation during track updates. Similarly, Deep OC-SORT [106] enhances OC-SORT [16] with appearance embeddings and camera motion compensation. SMILEtrack [176] further refines BYTETrack by improved appearance learning using transformer-based self- and cross-attention blocks. GHOST [152] follows a similarly lightweight philosophy, combining constant velocity prediction with simple Re-Identification (ReID) features. The authors use batch normalization for online domain adaptation for their separate ReID network and distinguishes between active and inactive tracks by aggregating appearance cues differently. This adapted cost matrix is then used with the Hungarian algorithm for robust association. StrongSORT [41] builds on DeepSORT [187] by introducing a set of enhancements. It adopts a stronger detector (YOLOX), integrates ReID features using a BoT-style network [101], and adds camera motion compensation. Further refinements include adaptive Kalman filtering and exponential moving average (EMA) updates for track features. StrongSORT++ extends this baseline further by introducing AFLink, a learned module for linking short tracklets, and GSI, a Gaussian smoothing interpolation method to fill detection gaps. Graph-based models like STTracking [191] and Relation3DMOT [22] formulate association via spatio-temporal or relational graphs using 2D and 3D features. Both methods incorporate appearance information into the graph construction to improve robustness in association. Further work include mask-guided networks [212], which extract region-specific features using instance segmentation, while occlusion-aware methods such as Hu et al. [66] separate short-term and long-term feature updates to prevent identity drift during occlusions. Lastly, works like [65, 66] incorporate quasi-dense similarity learning [130] and occlusion-aware appearance modeling.

Joint Detection and Embedding (JDE) methods deserve special mention as they aim to unify object detection and ReID feature extraction within a single network, rather than computing them separately. Examples include JDE [186], FairMOT [210], and CS-MOT [62]. While these approaches offer the advantage of efficiency by sharing a common backbone for both tasks, this architectural unification introduces certain trade-offs. In particular, the shared representation may limit the discriminative power of the appearance embeddings, making it harder to distinguish between visually similar objects compared to two-stage pipelines that use separate and specialized ReID modules. To address these challenges, FairMOT [210] proposes center-based representations along with

multi-scale feature fusion to better balance detection quality and embedding precision. Building on this, CS-MOT [62] introduces two architectural enhancements: an angle-center loss function that improves the training of appearance features, and a coordinate attention module that enhances spatial feature encoding.

Some works address unsupervised [94] or occlusion-aware [95] learning for improved ReID performance. Lastly, TripletTrack [109] uses an LSTM-encoded spatial history which is combined with learned appearance features to compute the affinity matrix used during data association.

The combination of motion and appearance continues to yield strong performance throughout the literature. JDE approaches, which unify detection and embedding tasks in a single network, offer compact and efficient pipelines. However, they may underperform in terms of embedding quality compared to dedicated ReID models. Additionally, many recent methods explore occlusion-aware association and the use of dynamic or long-term feature updates, which have proven important for maintaining identity consistency in challenging real-world environments.

TRACKING-BY-REGRESSION

In contrast to tracking-by-detection, tracking-by-regression approaches aim to directly predict object trajectories or displacements across frames, often eliminating the explicit data association step. These methods regress the position of objects from one frame to the next, typically by conditioning on the previous state or previous frame.

Tracktor [6] was one of the first works to demonstrate the viability of this paradigm in the MOT context. It builds on a two-stage object detector (Faster R-CNN [144]), reusing the regression head to predict object displacements in subsequent frames. This formulation eliminates the need for separate tracking modules, and can be enhanced by simple heuristics such as re-identification features or motion models. Other notable works include the GOTURN model [58], which was an early CNN-based single object tracker that regressed object locations based on appearance similarity between consecutive frames. While it does not scale to multiple objects or crowded scenes, it laid important groundwork for regression-based tracking.

CenterTrack [217] introduced a regression-based formulation where object centers are detected in the current frame, and their displacement to the previous frame is regressed jointly. This enables simultaneous detection and tracking with a shared representation. Gu et al. [50] extend this model to 3D tracking by incorporating depth and temporal attention features.

Overall, regression-based tracking models are efficient and avoid the combinatorial complexity of explicit matching, but they are generally less robust in the presence of occlusions or missed detections. Consequently, many modern regression-style trackers have incorporated hybrid mechanisms or transitioned toward transformer-based designs.

TRACKING-BY-ATTENTION

Transformer-based tracking approaches have recently emerged as a powerful and flexible alternative to traditional methods. These methods leverage attention mechanisms to model temporal dependencies and spatial relations across frames, often using query-based detection and tracking.

DETR-style architectures form the backbone of many modern transformer-based trackers. DETR (DEtection TRansformer) [17] is a transformer-based object detection framework that replaces traditional components such as region proposal networks and anchor boxes with a global set-based matching between object queries and predicted outputs. The core innovation of DETR lies in its use of learnable queries that attend to image features. These queries have proven especially valuable for tracking tasks, where maintaining object identity over time is crucial. As a result, DETRs end-to-end formulation and attention-based design have inspired a range of tracking extensions that build upon its representational strengths.

TransTrack [165] builds upon the DETR framework by introducing two parallel decoder branches: one for detection queries, which locate new objects, and one for propagation queries, which maintain object identities over time. This setup enables temporal continuity by explicitly modeling object propagation across frames. TrackFormer [110] refines this idea by combining detection and tracking queries into a shared decoder, which jointly reasons about current and past object locations. Although it still applies NMS to resolve prediction overlaps, TrackFormer improves data association by leveraging transformer attention across frames. MOTR [203] further advances this line of work by building on Deformable DETR [220] and removing the need for NMS entirely. It achieves this through the introduction of persistent track queries that span the entire video sequence, allowing the model to maintain consistent object identities without relying on post association.

Subsequent works introduce enhancements for occlusion handling and query assignment. DeNoisingMOT [45] uses denoising queries during training to improve robustness under occlusion. Yan et al. [192] propose one-to-set matching and more flexible assignment between detection and tracking queries. Memory-based architectures like MeMOT [15] maintain long-term appearance representations in external memory banks.

Cross-modal attention strategies are also being explored. MMFTrack [89] and Zhang et al. [208] combine LiDAR and image inputs via cross-modal attention. Boragule et al. [10] enhance deformable DETR [220] with pixel-level matching for stronger affinity estimation.

These transformer-based methods offer end-to-end tracking with flexible architectures, but they come at higher computational cost. DETR-based approaches have matured rapidly, whereas non-DETR alternatives often emphasize cross-modal fusion or long-term reasoning. The line between detection, tracking, and re-identification becomes increasingly blurred, as these components are fused into unified architectures driven by attention mechanisms.

2.3. DATASET GENERATION

This section discusses approaches that can be used to generate temporal Ground-Truth (GT), i.e., providing a temporal consistent ID and annotation (2D or 3D bounding box) for each object. The *de facto* standard of the last decades was the manual creation by human annotators at a dense frame rate (cf. [20, 47, 107, 164]). Early public labeling tools, like VIPER [111] or Labelme [202], provide standardized interfaces and label formats for this task. Both tools feature frame-wise annotation of arbitrary objects using

polygons, typically in the form of bounding boxes. Due to the labor-intensive nature of dense video labeling, and the often small inter-frame variation, keyframe annotation strategies became common. Here, labels are provided only at selected frames, and intermediate annotations are interpolated. A basic implementation assumes linear object motion between keyframes, enabling intermediate positions to be estimated with simple interpolation techniques, as seen in datasets like nuScenes [14]. More advanced interpolation schemes were adopted in datasets like the Caltech Pedestrian dataset [38], where cubic interpolation yields smoother trajectories [39]. Additional annotations are made based on the quality of the interpolation results. Further refinement is introduced by [48], who addresses the fact that constant motions of objects in the world do not project to constant motions in the image plane. Their method performs a cubic interpolation of the objects position in 3D. The image plane positions are used to solve a 3D reconstruction, which results are interpolated and back-projected into the image plane.

Beyond interpolation-based strategies, several early approaches explore semi-automatic annotation techniques to further reduce manual effort. VATIC [173] is one such system that uses bounding box information from multiple annotated keyframes to extract image features (color histograms and HOG) for each object. These features are used to train a discriminative classifier, which is applied to the intermediate unlabeled frames. Subsequently a dynamic programming approach is employed to calculate the final trajectories. The system is further extended in [172], by an active learning task to prioritize which keyframes and objects should be labeled next by a human annotator, thus improving annotation efficiency. In a different domain [119] demonstrated semi-automatic trajectory generation for one person captured in an indoor multi-camera setup. Here, the results of an ensemble of visual person trackers are fused and compared to the output of a reliable person detector extended by a human verification for uncertain estimations.

Complementary to these online or keyframe-based approaches, another line of work focused on offline, batch-based optimization methods for trajectory reconstruction. These methods process entire sequences jointly rather than frame by frame, enabling more global reasoning about object motion and identity. In [205], the data association problem for multi-object tracking is formulated as a flow problem and solved for the optimal set of correspondences by a min-cost flow algorithm. To account for occlusion the authors augment the network to include an explicit occlusion model. Following the graph-based formulation, Pirsavash et al. [135] introduce a greedy successive shortest-path algorithm with a greatly reduced runtime. More sophisticated formulations have been proposed by Andriyenko et al. [3], who formulates multi-target tracking as a discrete-continuous optimization problem. Thereby the data association is solved by using discrete optimization with label costs, whereas a continuous fitting problem is utilized to approximate the individual object trajectories. The discrete optimization is omitted in [113], where a non-convex continuous energy function is designed and minimized to solve the tracking task. The energy function combines partial image modalities in the form of appearance models and explicit object occlusion reasoning, with physical constraints, such as track persistence, mutual exclusion, and target dynamics. To find a strong local minimum of the non-convex function a combination of discrete jump-moves and continuous gradient descent is utilized. These early techniques laid the

groundwork for many of the more recent hybrid approaches and introduced key ideas for combining detection, appearance modeling, and temporal smoothing. An overview of more recent methods is shown in Table 2.1. The listed approaches are characterized by their employed detection method and modality, how temporal information is processed (object linking and filtering), the generated output, and the degree of required human annotation. Temporal processing is further broken down into tracker type, the similarity measure used for association, the motion model, the usage of ego-motion information, and if information from the keyframes is used during tracking. The tracker type describes what method is used to aggregate individual measurements into final tracks. For the similarity measure, distinctions are made between 2D (image) and 3D information used for spatial arrangement and appearance, respectively. The motion models define how objects are predicted between timestamps. Each approach is also evaluated for its reliance on ego-motion information. The generated ground-truth is categorized by its modality, i.e., 2D in the image or 3D, and the specified quality. The final column in the table reports the required level of human annotations.

Table 2.1: Overview of related work used for pseudo ground-truth generation for tracks. Methods differ in their employed detection method, how the resulting detections are linked/tracked over time. The amount and type of human supervision, and the generated GT results (2D or 3D). 2D refers to image-based coordinates. The constant-velocity motion model is abbreviated by CV. (?) indicates no mention in the paper, (-) indicates not applicable.

Name	Det.	Temporal Processing					Generated GT		Human Annot. Cost
		Tracker Type	Similarity Measure	Motion Model	Ego-Motion	Uses Key-Frames	2D/3D	Quality	
Offboard [136]	3D	Kalman Filter	Spatial (3D)	CV	Yes	No	3D	56.9AP ¹	69% 3D Annot. ²
Auto4D [193]	3D	Kalman Filter	Spatial (3D)	CV	Yes	No	3D	0.55 Recall ³	85% 3D Annot. ⁴
PePScenes [197]	2D	No tracking - temporal identity given by interpolated 3D annotations.					2D, 3D	?	2 Hz 3D Annot. ⁷ + manual corr.
MARS [214]	2D	Graph-based	Spatial (2D), Appearance (2D)	CV	- ⁵	No	2D	?	None
CityWalks [163]	2D	Kalman-Filter	Spatial (2D), Appearance (2D)	CV	No	No	2D	?	None
PathTrack [108]	2D	Graph-based	Spatial (2D), Opt. Flow	Opt. Flow	No	No	2D	0.9 Recall ⁶	Path Supervision
STIP [91]	2D	JPDA Filter	Spatial (2D), Appearance (2D)	CV	?	Yes ⁸	2D	?	2 Hz 2D Annot.
This thesis	2D	Three-stage	Spatial (2D), Appearance (2D)	CV	Yes	Yes	2D, 3D	66.1 HOTA	$\frac{1}{2}$ Hz 2D Annot.

¹ Compared to the performance of three labelers with 60.5 AP on five sequences of the WOD dataset [164]. Using an IoU threshold of 0.8.

² Trained on the full training subset of the WOD dataset [164]. Note that once the model is trained, it can be applied to more unlabeled data.

³ Numbers as stated in [193], using a manually labeled test set. IoU threshold of 0.8 was chosen for matching.

⁴ Trained on the full training subset of the Auto4D dataset [193]. Note that once the model is trained, it can be applied to more unlabeled data.

⁵ Not applicable, since surveillance dataset.

⁶ Numbers as stated in [108]. Tested on manually labeled test set. Using an IoU threshold of 0.5 for matching.

⁷ The annotations from the nuScenes [14] dataset are used, interpolated, and subsequently extended by the 2D detections.

⁸ Constrains generated tracks to 2 Hz GT. No additional information in the paper.

From Table 2.1, one observes a number of approaches [136, 193, 197] that involve 3D annotations. [136, 193] require a sizeable initial manually annotated dataset. [197] uses a high annotation frequency and an additional post-processing stage with manual corrections, increasing the annotation cost. Among the approaches using 2D annotations, some employ a forward-filtering paradigm [91, 163] used in online applications. Others

are graph-based approaches [108, 214] and operate in a batch-wise manner utilizing the offline nature to use information from the past and future.

Few works [108, 136, 193] touch upon the quality of the generated pseudo GT, and if so, only in detection context (i.e. without considering temporal correspondence). Furthermore, no quantitative analysis is provided on the trade-off between annotation cost and GT quality.

2.4. TRACKING DATASETS IN AD

Table 2.2: Overview of current datasets used for person tracking and motion prediction in automotive context. Object annotations and unique objects do refer to pedestrians and riders only. (?) indicates that the information was missing in the corresponding publication. Datasets are listed in ascending order of number of unique objects.

	#Frames (annotated)	# Object Annotations	# Unique Objects	Duration (min)	Geographical Coverage	Annotations	Annotation Procedure
KITTI-T [47] ¹	7.9 k	14 k	0.2 k	~ 13	Karlsruhe	2D BB, 3D BB, Segmentation	Manual Annot. [10 Hz]
View-of-Delft [128]	8.7 k	38 k	0.6 k	18	Delft (Netherlands)	2D BB, 3D BB	Manual Annot. [10 Hz]
MOT17 [112]	11 k	293k	1.3 k	7.7	? (9 scenes)	2D BB	Manual Annot. [30 Hz] ²
Argoverse [20]	350 k	143 k	1.5 k	60	Miami, Pittsburgh	3D BB	Manual Annot. [10 Hz]
PIE [139]	294 k	739 k	1.8 k	360 ³	Toronto	2D BB, Crossing, Attributes	Not specified [30 Hz]
JAAD [140]	75 k	391 k	2.8 k	46	5 Cities (4 Countries)	2D BB, Crossing, Attributes	Not specified [30 Hz]
nuScenes [14]	40 k	246 k	4.3 k	330	Boston, Singapore	3D BB	Manual Annot. [2 Hz] Interpolation [20 Hz] ⁴
ROAD [158]	122 k	295 k	5.0 k	170	Oxford	2D BB, Actions, Attributes	Manual Annot. [12 Hz]
Euro-PVI [151]	83 k	299 k	7.8 k	140	Brussels, Leuven	2D Segmentation 3D Locations	Manual Annot. [10 Hz]
TTTAN [107]	75 k	499 k	<14.0 k	175	Tokyo	2D BB, Attributes, Actions, Crossing	Manual Annot. [10 Hz]
STIP [91]	1108 k	3500 k	25.0 k	923	8 Cities (2 US states)	2D BB, Crossing	Manual Annot. [2 Hz] Semi-Supervised [20 Hz]
BDD100k [200]	400 k	509 k	27.5 k	1333	4 Areas (USA)	2D BB	Manual Annot. [5 Hz]
WOD [164] ⁵	401 k	(3D) 2863 k (2D) 2237 k	(3D) 23.9 k (2D) 46.6 k	676	6 Cities (USA)	2D BB, 3D BB	Manual Annot. [10 Hz]
ECP2.0	2070 k	9982 k	228 k	1725	29 EU Cities (11 Countries)	2D BB, 3D Locations	Manual Annot. [0.5 Hz] Semi-Supervised [20 Hz]

¹ Statistics have been generated on the public train subset. Missing the 29 test sessions.

² Sequences have different image frequencies (14 Hz, 25 Hz, and 30 Hz). Annotations have been performed for full frequency.

³ Duration stated as data recorded in total.

⁴ Manual annotations are interpolated to full frequency for each sensor (camera 12 Hz, LiDAR 20 Hz).

⁵ Data available with the June 2022 update used to generate the statistics.

Table 2.2 provides an overview of datasets for the task of person tracking and motion prediction in automotive context. It only lists datasets which include image data. Recent automotive datasets for 3D object detection use LiDAR-based annotations [14, 20, 128, 164]. Aside from [128] which focuses on radar applications, the datasets offer a significantly higher number of objects compared to previous work [47]. Annotations are provided in the form of 3D bounding boxes [14, 20], while [164] and [128] also offer corresponding 2D information. The image-based BDD100k [200] comes at a comparable size, containing solely 2D annotations. Other tracking datasets outside the automotive

context have been published. First, the STCCrowd [32] dataset focuses on person crowds and offers 2D and 3D annotations for synchronized LiDAR and image data. Second, the PersonPath22 [154] dataset contains images mainly captured by static cameras with 2D bounding box annotations.

In contrast, intention prediction focused datasets offer additional attributes (e.g., action, attributes, crossing indication) accompanying the 2D bounding boxes [91, 107, 139, 140, 158, 163], or 3D annotations [151, 197]. Most datasets have been recorded in one or few different cities, offering only limited geographical coverage. Notable exceptions are STIP [91] recorded in six different cities in two different U.S. states, JAAD [140] with five cities in four different countries, and lastly CityWalks [163]. The latter includes recordings from 21 different European cities, by using non-automotive videos from YouTube. The last row of the table lists the ECP2.0 dataset presented in this thesis. A detailed description of its creation and characteristics is given in Chapter 4.

2.5. PERSON RE-IDENTIFICATION

This section reviews recent advancements in person ReID, with a focus on deep learning applications and the integration of multi-modal data. For foundational methods, readers may refer to the comprehensive survey by Zheng et al. [213]. Current strategies predominantly leverage deep learning to extract unique person features through positive and negative sample training, often integrating triplet loss [59] with cross-entropy loss [215] for metric learning.

CNN-based methods have been extensively explored, including notable approaches such as BoT [101], SBS [56], AGW [198], and MGN [175]. BoT [101] set a new baseline by incorporating a series of enhancements into the training strategy, which is further improved by SBS [56]. AGW [198] extends BoT by integrating attention mechanisms into the CNN architecture, while MGN [175] extends the CNN to a multi-branch network structure to simultaneously learn global and local features. Recent research has increasingly focused on addressing occlusions, a major challenge in the field, which introduce non-object appearances into the learning process. To mitigate this, data augmentation-based approaches [25] train networks to recognize and disregard occluded regions by artificially occluding parts of the image during training. Other approaches segment the person patch into several parts and learn local part-based representations [160, 181, 196, 216]. These parts can be created using semantic networks [196], information from a pose estimation network [160, 216], or predefined static areas such as a fixed set of rows [181]. Part-based approaches generate a feature embedding for each part individually, while also extracting visibility [160, 196, 216] or quality [181] scores. Other techniques address occlusion by synthesizing complete images from sequential data [63], using human shape as additional supervision [218], or by excluding occluded samples during training [95].

Transformer-based methods have recently gained traction in person ReID, with models such as TransReID [57], AAFFormer [219], Performer [133], DC-Former [87], LoGoViT [134] adapting the Vision Transformer (ViT) architecture [40] to this task. TransReID

[57] presents the first ViT-based ReID framework incorporating enhancements such as overlapping patches and embeddings that capture orientation or camera-specific information. Various extensions have been proposed [57, 87, 134, 219] to improve the models performance and its ability to handle partially occluded persons. TransReID [57] utilizes patch tokens in different permutations in addition to the class token, while LoGoViT [134] processes the patch tokens through an additional transformer layer. DC-Former [87] increases the number of class tokens, whereas AAFomer [219] introduces new part tokens.

Multi-modal and LiDAR-only methods. The integration of LiDAR or depth data into ReID is relatively uncharted. Early techniques employed RGB-D data (image and depth) to derive anthropometric features [90, 92, 116], or utilized skeleton data from the Kinect RGB-D sensor [131, 188]. More recent transformer-based approaches have merged RGB with depth data [21, 115] or other auxiliary inputs like depth or contour plots [182]. In addition to depth data, a body of work has focused on multi-modal ReID using RGB and thermal imagery [185, 194, 206]. These modalities can be processed separately [185, 194] or jointly by a common backbone [206]. Information across modalities is shared using token permutation [185], spatial- and frequency-based token selection and aggregation [206], or an unsupervised collaborative learning strategy utilizing deep and shallow features [194]. Recently, two methods have been presented which focus on solely using LiDAR points for re-identification [51, 168]. [51] presents a graph-based complementary enhancement encoder to extract features from multiple pointclouds followed by a transformer-based temporal fusion to estimate the final ReID features. In [168] a Siamese network tracker is extended to generate ReID features for vehicles and pedestrians using the pointcloud input.

2.6. DISCUSSION

This chapter reviews prior work in visual multi-object tracking and person ReID, covering traditional tracking pipelines (Section 2.1), deep learning-based approaches (Section 2.2), dataset generation strategies (Section 2.3), tracking benchmarks in automated driving (Section 2.4), and person re-identification methods (Section 2.5). The survey highlights a clear methodological transition from modular model-driven pipelines, toward increasingly integrated, data-driven, and end-to-end learning-based approaches.

While demonstrating strong performance, many state-of-the-art tracking approaches implicitly rely on dense and reliable supervision, yet existing dataset are often constrained in scope and temporal or geographical coverage. This is particular evident in complex urban scenes characterized by frequent occlusions, close interactions between Vulnerable Road Users (VRUs), and long-tailed behavior, which are insufficiently represented in many existing tracking datasets. The resulting mismatch between data requirements and available annotations represents a persistent bottleneck for further progress.

This chapter further underscores the importance of robust identity association in crowded traffic environments, as discussed in Section 2.2 and 2.5. Motion cues alone are frequently insufficient over longer time horizons, motivating the use of appearance embeddings and person ReID techniques.

Overall, this chapter suggests that continued advances in visual tracking for auto-

mated driving are not driven by algorithmic innovation alone, but are fundamentally constrained by the availability of suitable tracking data. Reducing annotation effort while preserving temporal and identity consistency emerged as a key prerequisite for scalable learning-based tracking, directly motivating the focus of the subsequent chapters of this thesis.

3

SEMI-SUPERVISED PIPELINE FOR CREATING A PERSON TRACKING DATASET

The previous chapter highlights the growing need of modern tracking approaches for large-scale annotated data, which remains costly to obtain. This chapter introduces a semi-supervised trajectory generation pipeline that reduces manual annotation effort while producing temporally dense person trajectories. In addition to presenting the pipeline design, this chapter provides a systematic experimental evaluation of trajectory generation strategies, ablations, and limiting factors, establishing the selected approach that is used to generate the annotations of the dataset presented in Chapter 4.

3.1. PIPELINE OVERVIEW

In the following an overview of the proposed pipeline for the semi-supervised creation of an Automated Driving (AD) dataset is given. The approach operates on data streams of images, LiDAR point clouds, and ego-motion information. In addition, the sparse supervision of annotated keyframes is used, in the form of 2D bounding boxes in the image with temporal correspondence. For all the intermediate (non-keyframe) images, which have not been manually annotated, an automatic detection method is employed, generating a set of person bounding boxes (cf. Section 3.2.1). As both annotations and detections are represented as 2D bounding boxes in the image, the combined set is referred to as *2D locations*. The generated and annotated person bounding boxes are uplifted to 3D using the point cloud information, which is further detailed in Section 3.2.2. As the 3D information of the LiDAR is relative to the ego-vehicle, the uplifted position is referred to as *relative 3D location* (not limited to any ground-plane). The individual relative 3D positions are compensated by the ego-motion of the vehicle using a flat-world assumption (i.e., vehicle pitch and roll are not taken into account, see Section 3.2.3). This results in a world-fixed position for each bounding box, called *3D locations* in the following.

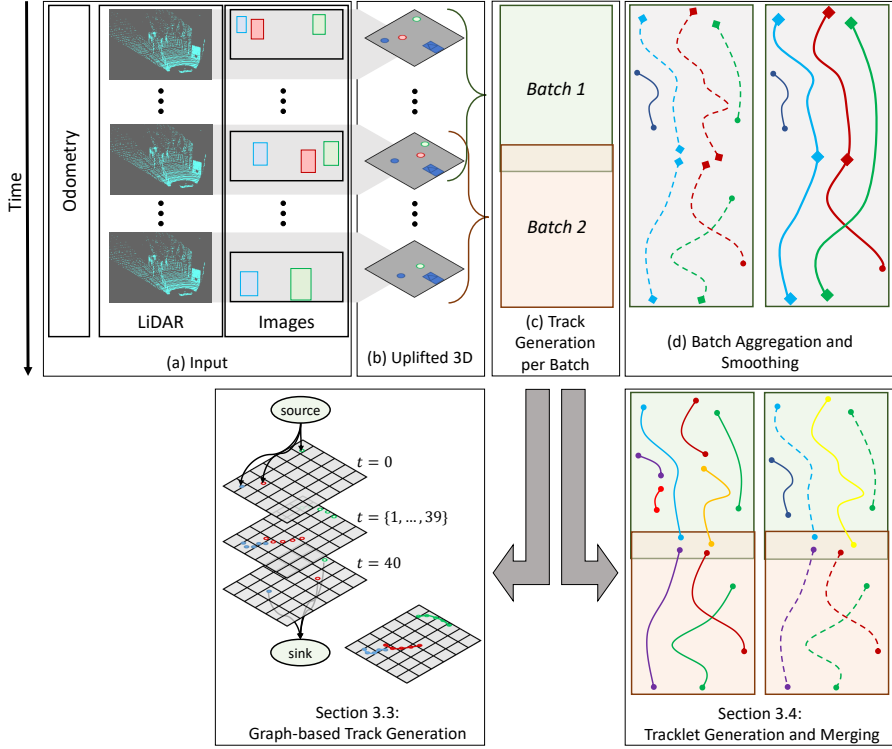


Figure 3.1: Schematic illustration of the proposed pipeline to generate dense person trajectories from sparse image annotations and unlabeled additional input. (a) Inputs include sparse manual 2D and identity annotations at keyframes (three of which shown) and unlabeled intermediate images, LiDAR point cloud, and odometry data. Based on these inputs, an uplifting to 3D is performed, for which the results are depicted in (b). (c) For each batch, starting and ending with a keyframe, tracks are created either using the graph-based optimization approach (Section 3.3, left) or the tracklet-based method (Section 3.4, right). (d) The merged tracks for each batch are aggregated (dashed tracks) (left) using the identity information available in the keyframes, and are finally smoothed (right). Diamonds indicate the use of keyframe annotation.

Lastly, the combination of all aforementioned information for one timestamp is called *measurement* (distinguished into *annotation* or *detection* based).

At this point, temporal correspondence is to be established. The data is separated into batches, such that each batch starts and ends with a keyframe. This thesis presents two different sub-approaches to generate tracks for each batch. First, a constrained flow optimization method is employed, which is further detailed in Section 3.3. All 3D locations within a batch are reformulated as a spatio-temporal Directed Acyclic Graph (DAG). Tracks are generated by extracting the k shortest paths in the graph. Second, Section 3.4 presents a two-stage tracklet-based tracking approach. In the first stage, image measurements (location and appearance) are employed to generate a set of image-based *tracklets* within a batch. 2D tracklets are generated to reliably link individual object measurements, without enforcing to capture the whole trajectory of an object. To combine the tracklets into *tracks*, which possibly extend to the entire batch, a merging method

is introduced that takes the accumulated 2D appearance and location of the measurements into account.

The object identity information from the keyframes is subsequently used to aggregate the per-batch tracks into *trajectories*, i.e. that extend across batches. Lastly, the resulting trajectories are smoothed using a final post-processing, which uses all given measurements. The output is a set of object trajectories in 2D and 3D. Figure 3.1 gives an overview of the whole semi-supervised pseudo Ground-Truth (GT) generation pipeline.

The following subsections provide more details regarding the various steps.

3.2. DATA PREPARATION AND PREPROCESSING

3.2.1. DETECTION AND APPEARANCE EMBEDDINGS GENERATION

Detections for the non-keyframes are generated using the Pedestron detection network [54], the top-ranking publicly available method on the ECP1.0 detection benchmark. Given each 2D location (both annotated and generated) a pre-trained appearance embedding network is employed. For each person measurement given by a bounding box in the image, a corresponding image patch is extracted. Hereby, this thesis utilizes and extends an approach originally presented for person re-identification [59]. The patch is used as the input for the network, which acts as embedding function to transform the image information into a latent appearance space. This transformation is learned in a way, such that patches showing the same person result in close locations in the latent space, i.e., the distance between them is small. ResNet-50 [55] is employed as base network in conjunction with a multi-task loss design including the Ranked List Loss [184] and identity classification loss. Following [101] several extensions to the network and the training procedure are introduced. Namely, warm-up learning rate, random erasing augmentation, label smoothing, batch normalization, and a reduction of the stride of the last convolutional layer. More information is given in Appendix 3.A.

3.2.2. 3D UPLIFTING

The goal is to get the relative 3D location of a 2D bounding box annotation or detection in the camera coordinate system. To this end the ego-motion compensated and projected LiDAR points are used to identify point clusters representing the annotated person. The exact details are as follows.

For each person annotation defined by its bounding box $(u_1^*, v_1^*, u_2^*, v_2^*)$: first, an initial point set is created containing all LiDAR points p_i^l with a corresponding projection p_i inside the bounding box. Where $i < n$ and n is the number of all LiDAR points within a 360° scan. Second, the height distribution of persons serves as prior knowledge, allowing verification of whether a given point p^c could have been reflected/caused by a person. Using a pinhole camera model, the relationship between an objects height in pixels $\Delta \tilde{v}$ and its real height Δy^c is given by

$$\frac{\Delta \tilde{v}}{\tilde{f}} = \frac{\Delta y^c}{z^c}, \quad (3.1)$$

with $\tilde{f}$ being the focal length and z^c the z coordinate. The assumption is made that the object is standing upright and the 3D points defining upper and lower boundary of

the 2D box share the same z^c coordinate. Given the distance of the point z^c and the pixel height $\Delta \tilde{v}$ of the rectified bounding box, the potential object height Δy^c in meters is computed according to Eq. (3.1). This height is expected to be in the range of $[\Delta y_{min}^c, \Delta y_{max}^c]$, with $\Delta y_{min}^c = 0.7 \text{ m}$ and $\Delta y_{max}^c = 2.0 \text{ m}$. The lower bound is given by the 0.1% quantile of the height distribution of 18 months old girls [122], while the upper bound is given by the 99.9% quantile of the height distribution of adult men [171]. All points resulting in object heights outside of this range are expected to be caused by noise (e.g.: background or occlusions), and are thus removed from the person's points set.

Third, using a scene segmentation method [24] trained on CityScapes [33], the predicted class at the projected image position p_i is assigned to each LiDAR point. All LiDAR points which do not support the bounding box class (pedestrian or rider) are excluded from the set of points associated to the current bounding box.

Finally, point clusters are extracted from the filtered point sets. Objects with overlapping bounding boxes are grouped together. For each group the filtered point sets are united and processed jointly, to prevent assigning one point to multiple objects. An extended version of the constrained k-means clustering algorithm [174] is employed to find k point cluster for the point set of every object group, with k being the size of the group. In [174] the original k-means clustering is extended by a set of *must link* and *cannot link* point tuples. For the approach presented in this thesis solely the cannot link set is used to prevent that points which cannot belong to the same object - as they lie within different bounding boxes - end up in the same cluster. Furthermore, points with a distance greater than 1.5 m cannot be caused by the same person and are thus also added to the cannot link set. In the original version the algorithm fails as soon as a point cannot be assigned to any cluster. To avoid that, the algorithm is extended by introducing an outlier class for all points without a fitting cluster. The number of initial clusters is set to l , generation of additional clusters is allowed during the first iterations of the algorithm (to account for additional depth modes e.g. caused by occlusions, those points have not been filtered by the previous steps). For the final cluster to object assignment, the objects are sorted by their occlusion level if available and base point v_2^* in image coordinates. Thus, non-occluded, closer objects - assuming a flat world - are processed first. For each object the cluster with the highest number of points within its bounding box is chosen. Clusters are only assigned to objects once. The final 3D position is defined by the intersection point of the ray of sight through the 2D bounding box center with the x-y-plane at the mean distance $\bar{z}^c$ of all points within the assigned cluster.

If not mentioned otherwise, the distance is always defined by the distance value along the z-axis, which is orthogonal to the image plane. For an overview of the label lifting see Fig. 3.2.

3.2.3. EGO-MOTION COMPENSATION

To perform the ego-motion compensation the internal car sensors are utilized. The vehicles wheel encoders measure the rotational speed of each wheel. By using the information from the wheel speed sensors, the known axle configurations, along with an Ackermann steering model, the current velocity in the heading direction as well as the yaw rate are calculated.

Between two time frames the ego-motion of the vehicle is calculated in a XY driving

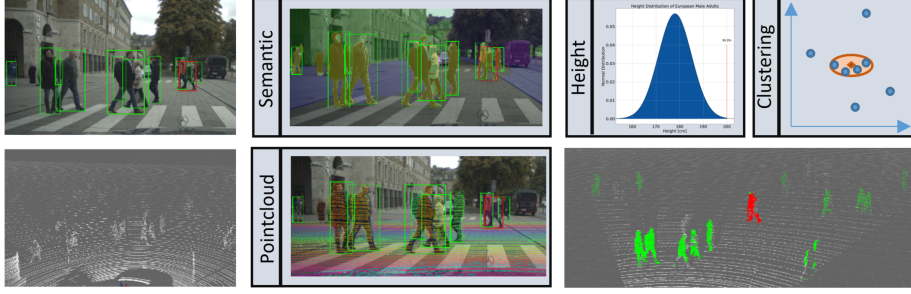


Figure 3.2: Overview of the 3D lifting procedure. Inputs are 2D bounding boxes and 3D point clouds (left column). The LiDAR points are ego-motion corrected and projected into the image (middle column bottom). For every bounding box only points within the bounding box supported by the scene segmentation (middle column top) are potential candidates. After applying additional constraints the final point clusters per bounding box are found using a clustering step (right).

plane, which is assumed to be fixed in between tracking cycles. To calculate the change in position (X and Y) of the car, the velocity v in driving direction, the yaw rate $\dot{\varphi}$ (rotation around Z axis), and the time delta $\Delta t = t - t_{-1}$ need to be known. By assuming a fixed velocity and yaw rate, the change in angle and position can be calculated by:

$$\Delta\varphi = \dot{\varphi} \cdot \Delta t \quad (3.2)$$

$$\Delta x = \frac{v}{\dot{\varphi}} \cdot \sin(\Delta\varphi) \quad (3.3)$$

$$\Delta y = \frac{v}{\dot{\varphi}} \cdot (1 - \cos(\Delta\varphi)). \quad (3.4)$$

The change of the vehicle position and angle can be expressed as the affine transformation

$$T_{ego} = \begin{bmatrix} \cos(\Delta\varphi) & -\sin(\Delta\varphi) & \Delta x \\ -\sin(\Delta\varphi) & \cos(\Delta\varphi) & \Delta y \\ 0 & 0 & 1 \end{bmatrix}. \quad (3.5)$$

Finally the compensated X and Y position of each object can be calculated by:

$$\begin{bmatrix} x_{comp} \\ y_{comp} \\ 1 \end{bmatrix} = T_{ego}^{-1} \cdot \begin{bmatrix} x_{old} \\ y_{old} \\ 1 \end{bmatrix}. \quad (3.6)$$

The ego-motion compensation is depicted in Figure 3.3.

3.3. GRAPH-BASED TRACK CREATION

By transforming all ground-truth annotations and detections into a global world frame and correcting them by the ego-motion of the vehicle - as described in the previous Section - for each time step there is a set of 3D object locations. To find the correct correspondences between these positions, which are needed to extract the final object tracks,

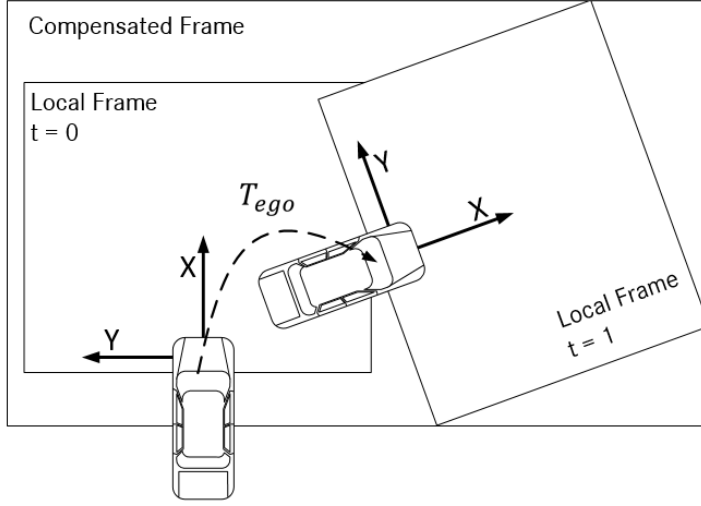


Figure 3.3: Schematic depiction of the ego-motion compensation. At each time, the measurements are received in a local coordinate system. By applying the inverse of the ego movement, the measurements can be reported in the previous coordinate frame.

the k-shortest path (KSP) approach presented by Berclaz et al. [5] is adopted, including the extension by Shitrit et al. [153], which introduces so-called identity groups to disambiguate persons based on their appearance.

The algorithm operates by discretization both in the temporal and spatial domain. As input for each time-step a Probabilistic Occupancy Grid Map (POM) is used. The POM is a two-dimensional representation of the XY world plane, with a given number of equally spaced cells in X and Y. Each cell holds the probability of the presence of an object at this location. The final representation of a sequence is a three-dimensional graph (two spatial dimensions, and the time), where the nodes indicate the position and presence of an object, and the edges indicate object movement. The optimization is performed in a batch wise manner. Thereby a batch is constructed between two keyframes, and the detection responses in between those two frames are used as input for the path optimization.

To enable the usage of sparse appearance information the extension from Shitrit et al. [153] is integrated into the KSP. A set of L identity groups is introduced, which partition the set of tracked persons. The original graph is multiplied L times, introducing a separate layer per group. There are no edges between group layers, i.e., a path cannot represent two different identity groups. Within one layer, the edge costs are calculated by taking the existence probability from the POM and an additional appearance model for the current layer.

Based on the constructed graph, the path optimization works as follows: The set of all possible locations an object can reach at time $t + 1$ given its location k at t is denoted by $\mathcal{N}(k) \subset 1, \dots, K$. Each edge $e_{i,j}^t$ indicates a possible object movement from i at time t to j at time $t + 1$, implying that $j \in \mathcal{N}(i)$. Lastly, the actual number of objects moving

from i to j is represented by the discrete variable $f_{i,j}^t$.

Based on these definitions the following constraints are formulated. For each t , the sum of outgoing flows $f_{j,k}^t$ of a location is equal to the sum of incoming flows $f_{i,j}^{t-1}$ to this location at the previous time. Furthermore, a node cannot be occupied by more than one object at a time. Lastly, there can be no negative flows.

The appearance and disappearance of objects in the batch is enabled by a source and sink node. All flows need to emerge from the source node, and ultimately end up in the sink node. Given these definitions, and since object movement can only occur from $t-1$ to t (objects only move forward in time), the graph is directed and acyclic.

Thus, to find the set of correct correspondences between object locations over time for one object translates into finding the best path in the graph. Finding this solution can be expressed as an Integer Programming (IP) - maximizing the flows in the network - which could be solved by a generic Linear Programming solver. By leveraging the directed acyclic graph representation, the IP can be reformulated as a k shortest node-disjoint paths problem which significantly reduces the complexity. By using the iterative disjoint path algorithm [166] the k-shortest paths in between the source and sink node are extracted, which is the optimal solution to solve the assignment problem. An illustrated example is provided at the bottom left of Figure 3.1, depicting the constructed graph with its source and sink nodes, the aggregated positions in between, and final extracted paths.

3.4. TRACKLET-BASED TRACK CREATION

The tracklet-based method presented in this section first generates short-term tracklets (Section 3.4.1), which are subsequently merged into longer tracks (Section 3.4.2). An illustrative example is shown at the bottom right of Figure 3.1.

3.4.1. TRACKLET GENERATION

This thesis uses DeepSORT [187] as the online tracking method to link individual 2D measurements into tracklets. Tracking is performed in the image space, using a Kalman filter with a linear motion model. The state space $[u, v, a, h, \dot{u}, \dot{v}, \dot{a}, \dot{h}]$ contains the center location (u, v) , the aspect-ratio a of the objects bounding box, and the height h , along with their respective velocities. The association between active tracks and new measurements is performed in a two-step procedure involving appearance and position features.

In both steps, the Hungarian algorithm [77] is used to calculate the optimal associations given a cost matrix (costs between tracks and measurements). In the first step, the appearance information are used to calculate the cost matrix. For each measurement i , the appearance feature ϕ_i is obtained from the appearance descriptor network as described section 3.2.1, replacing the original appearance embedding method of DeepSORT. The appearance feature vectors ϕ_i and ϕ_j of two measurements are compared using the $L2$ norm, such that:

$$\Delta_{app} = \|\phi_i - \phi_j\|. \quad (3.7)$$

If the distance exceeds the pre-defined threshold γ_{app} , the pair is discarded (no association based on appearance information is possible). In the second association step, the 2D Intersection-over-Union (IoU) in the image space is used to calculate the cost matrix

for all tracklets and detections which have not been matched in the first stage. Again a pre-defined threshold γ_{iou} is used to prevent matches of pairs with a too small IoU value. All tracklets without an associated measurement after both cycles are marked as missed, whereas unmatched detections are used to create new tracklets. The associated measurements with $[u, v, a, h]$ are used to perform a Kalman-update of the tracked state space using a fixed measurement noise.

Track management follows the one introduced in [187], such that each tracklet has an associated tracklet status for each timestep. A newly created tracklet enters the *initialization* status. After n_{init} successful updates, the tracklet status is changed to *confirmed*. Otherwise, if a tracklet in the *initialization* status has no associated measurement in a timestep (miss), it is directly deleted. Confirmed tracklets are kept alive until a pre-defined number n_{missed} of consecutive misses is reached, after which they are deleted.

The original track management is extended, to leverage the GT information contained in keyframes. Tracklets created based on a keyframe annotation directly enter the *confirmed* tracklet status. This allows the tracklet generation to resolve some ambiguities as the reliable keyframe annotation can be used as starting point in the presence of unsteady detections. In addition, the bounding box information of the GT is used instead of the detections.

3.4.2. TRACKLET MERGING

For combining tracklets this thesis adapts and extends the Forest Path Cutting (FPC) algorithm [100]. The method was introduced to fuse segmentation proposals over time for unsupervised offline video segmentation of arbitrary objects. The general graph creation and optimization scheme is extended and the cost calculation is adapted to enable the integration of keyframe annotations.

The FPC algorithm consists of two steps. In the first step, the set of input tracklets is used to create the name-giving forest. Each tree in the forest represents a directed acyclic graph, representing possible tracklet associations. Each node in the graph represents a tracklet τ_i , while edges between two nodes represent a possible association of two tracklets with an associated cost. For the creation of the association forest, it is assumed that tracklets have at most one parent tracklet and that tracklets do not overlap, i.e., a tracklet for an object ends before another tracklet for the same object begins. Thus, to create the forest, the tracklets are ordered descending by their start time. The set of possible merge candidates $\mathcal{T}^i$ for a target tracklet τ_i is defined by all tracklets ending prior to the target tracklet, such that $t_i^{start} >= t_j^{end}$. For each tracklet τ_j in the candidate set $\mathcal{T}^i$ the association costs $\Delta_{i,j}^{tracklet}$ is calculated. The tracklet τ_j which minimizes this association cost is set as the parent tracklet for the target.

In [100] the association cost is calculated based on the normalized distance of the tracklets mean appearance vectors and a temporal density score. The cost calculation is tailored by introducing an additional term and removing the temporal density score, such that the new cost is given by:

$$\Delta_{i,j}^{tracklet} = \Delta_{i,j}^{app} + \lambda \Delta_{i,j}^{dist}, \quad (3.8)$$

where λ is a weighting factor and the two distance functions are given by

$$\Delta_{i,j}^{app} = \|\bar{\phi}_i - \bar{\phi}_j\|, \quad (3.9)$$

$$\Delta_{i,j}^{dist} = BC(\hat{\tau}_i, \hat{\tau}_j). \quad (3.10)$$

The mean appearance vector $\bar{\phi}$ is defined as the average of all appearance vectors of the bounding boxes of one tracklet. For the newly introduced spatial distance term $\Delta_{i,j}^{dist}$, the Bhattacharyya distance (see Eq. 3.11) is employed to compare the smoothed and predicted 2D locations of the tracklets. To account for measurement noise and errors in the online tracklet generation, both tracklets are smoothed using all available measurements using the interval-based Rauch-Tung-Striebel (RTS) algorithm [141]. Using the smoothed states (position and velocity) both tracklets are predicted to the same intermediate time t^{middle} for comparison. As described above, the candidate tracklet τ_j ends prior to τ_i , such that t^{middle} is given by $(t_i^{start} + t_j^{end})/2$. Prediction is thus performed forward in time for τ_j , while τ_i is predicted backwards.

This results in the predicted state and covariance for both tracklets given by $\hat{\tau}_i = \mathcal{N}(\mu_i, \Sigma_i)$ and $\hat{\tau}_j = \mathcal{N}(\mu_j, \Sigma_j)$, which are used to calculate the Bhattacharyya distance as in [82] with:

$$BC(\hat{\tau}_i, \hat{\tau}_j) = \frac{1}{8}(\mu_i - \mu_j)^T \hat{\Sigma}^{-1}(\mu_i - \mu_j) + \frac{1}{2} \ln \left(\frac{\det \hat{\Sigma}}{\sqrt{\det \Sigma_i \det \Sigma_j}} \right), \quad (3.11)$$

where $\hat{\Sigma} = \frac{1}{2}(\Sigma_i + \Sigma_j)$, and τ_i .

In the second step of the algorithm, the individual tracks are extracted. By applying the FPC the path with the smallest cost is selected from the forest, the nodes connected by this path are removed, and the forest is re-arranged accordingly. This is done in an iterative fashion until all paths are extracted, and thus no more tracklets can be merged.

Pre-defined validation gates are used to prevent the merging of dissimilar tracklets. Namely, if the spatial distance $\Delta_{i,j}^{dist}$ exceeds the threshold Γ_{dist} , or the mean appearance distance $\Delta_{i,j}^{app}$ exceeds the threshold Γ_{app} , no edge is added between the nodes.

Lastly, as identity information are available for keyframes, the algorithm was extended to be able to utilize this information. After creating the association forest, paths that contain different object identities based on the information from the keyframes are split. Splitting is performed, such that the costs of the resulting sub-paths are minimized (i.e., the path is split such that the connection with the highest association cost is removed). The sub-paths are then correctly added to the existing forest. Additionally, the matching cost of tracklets containing the same keyframe annotation identity is set to zero.

3.5. BATCH AGGREGATION AND TEMPORAL REFINEMENT

Given the batch-wise approach used to generate the tracks (cf. Section 3.3 and 3.4), the individual tracks need to be combined across the batches into trajectories. This is done by using the keyframe annotations. Each object in each keyframe holds a unique object ID. These identity information are preserved when using a keyframe annotation for the tracklet generation. The correspondence between two tracks in two different batches is thus given by their shared keyframe identity. The object might be occluded and thus not labeled in one of the keyframes. Still, linking is possible as long as the object is not occluded in the preceding and next keyframes. As the identity information is directly used, there are no ambiguities introduced by differences in the estimated position at the end of a batch compared to the starting position in the new batch. The only assignment errors that can arise are due to a faulty association during the tracklet generation.

To mitigate uncertainties both in the 2D locations as well as in the uplifted 3D locations, a final Kalman smoothing post-processing is employed for each generated object trajectory, by using the interval-based RTS algorithm [141]. Smoothing is performed both for the bounding boxes in the 2D image space, as well as for the 3D locations only taking the Birds-Eye View (BEV) coordinates into account (namely X and Y , by employing the flat-world assumption). For the 2D bounding box smoothing, the state space is designed identical as for the online tracker described in Section 3.4.1, with $[u, v, a, h, \dot{u}, \dot{v}, \dot{a}, \dot{h}]$. For the 3D smoothing, the state space is defined by the objects center position in the world-fixed BEV, its velocity, and its metric height with $[X, Y, \dot{X}, \dot{Y}, H]$. For both the 2D and 3D filter, a Constant Velocity (CV) motion model is used for the state transition matrices (except for the metric height H of the 3D smoother, which is assumed to be constant). In addition, fixed individual state process noise matrices Q^{2D} and Q^{3D} are used. The measurement noise of the 2D filter R^{2D} is assumed to be fixed, while the uncertainty calculated during the uplifting is used as the measurement noise R_i^{3D} for each measurement i during 3D smoothing.

For smoothing, only the raw measurements are used (not the already filtered track states). In the case of missing measurements for a timestep, the filter still generates an optimally estimated state using all information. Thus, missing measurements due to occlusion or detector errors can be compensated. The generated smoothed trajectories are clipped, such that they start and end with a measurement (no extrapolation).

Due to the keyframe annotations, the objects states for the keyframes are known. This information is incorporated into the smoother by adapting the measurement noise of the keyframe annotation input. As the 2D annotations in a keyframe have been created by human annotators, their uncertainty is an order of magnitude smaller than those of the detector. Thus, the RTS smoother adapts the gain accordingly during the state update, while taking the past and previous positions into account. This also translates into the 3D case, where the uncertainty along the ray-of-sight is kept.

3.6. EXPERIMENTAL VALIDATION

This section presents the experimental evaluation of the pseudo ground-truth generation pipeline. All experiments in this section are performed on the ECP2.0-dense validation dataset, which provides manually annotated ground-truth trajectories at an in-

creased temporal resolution and is described in detail in Section 4.6. The evaluation focuses on assessing the quality and robustness of the pseudo GT trajectories generated by the two alternative approaches introduced in this chapter. Performance is measured using the Higher Order Tracking Accuracy (HOTA) metric together with its associated sub-metrics capturing detection, localization, and association quality, as introduced in Section 4.5. This enables a balanced comparison of spatial accuracy and temporal identity consistency across methods.

The experimental investigation is organized into three parts. First, the proposed graph-based and tracklet-based approaches are compared against each other and against baseline methods. This initial comparison focuses on the overall quality of the generated trajectories and serves to identify the most suitable pseudo GT generation method among the considered alternatives. Based on these results, the tracklet-based pipeline is selected for all subsequent investigations.

Second, a series of ablation studies examines the impact of key design choices within the selected tracklet-based pipeline, including the incorporation of keyframe information, the contribution of the individual processing stages, and the effect of batch-wise aggregation and post-processing. These experiments provide insights into the relative importance of the different components.

Third, a ceiling study investigates the sensitivity of the selected tracklet-based pipeline to detection quality by replacing the noisy detector outputs with idealized detections. This analysis highlights the dominant error sources in the generation of pseudo GT trajectories.

3.6.1. COMPARISON OF PSEUDO GT GENERATION METHODS

As a first external baseline, linear interpolation of the bounding box annotations in the keyframes is used to generate pseudo GT. This includes interpolation in the image space (*interp-image*) for $[u, v, w, h]$. This has the drawback not to account for the vehicle ego-motion. Therefore, interpolation in 3D is used as second baseline, by interpolating the object's location on the ground plane $[X, Y]$ in the ego-motion compensated world-fixed BEV (*interp-world-ego*), and using the interpolated locations to project the object into the image.

The third baseline is the DeepSORT-based [187] approach used in [163] for the creation of person trajectories. As DeepSORT is also employed for tracklet generation in this thesis, the sequence-based DS variant (cf. first row in Table 3.2) is used for comparison. Following the strategy outlined in [163], all tracks shorter than 3 seconds are discarded.

The fourth evaluated method is the *graph-based* KSP (K-Shortest-Path) optimization method presented in Section 3.3. For the KSP approach, ground-truth data for the intermediate frames is used.

The *tracklet-based* method from Section 3.4 (cf. $DS^K + FPC^K$ in Section 3.6.2) is used as fifth and last method.

Table 3.1 shows the numbers for all five methods. Given the interpolation reliance on two keyframe annotations, a separate evaluation setting is defined, considering only ground-truth objects which exist in at least two adjoining keyframes, see top part of Table 3.1. This reduces the number of objects by roughly 50% (see Table 4.2). Results are reported both for all object tracks and for the subset of tracks with a minimum duration

Table 3.1: Assessment of pseudo ground-truth quality on the ECP2.0-dense validation dataset. The bold entries correspond to the pipeline introduced in this Chapter, using either *graph-based* or *tracklet-based* per-batch track generation. Interpolation was performed in the image or in 3D space. The top part only considers performance for ground-truth objects which appear in at least two adjoining keyframes. The middle part shows the results for all **GT** objects. The bottom part shows the evaluation results only considering tracks longer than three seconds.

Eval	Method	HOTA $\uparrow$	AssA $\uparrow$	DetA $\uparrow$	LocA $\uparrow$
≥ 2 adj. keyfr.	interp-image	41.1	49.4	34.5	78.8
	interp-world-ego	54.6	62.2	48.1	80.3
	graph-based	45.9	35.5	59.6	94.7
	tracklet-based	68.8	73.6	64.5	90.0
all	interp-image	34.8	42.2	29.0	78.6
	interp-world-ego	45.2	53.2	38.6	80.1
	graph-based	41.8	32.3	54.2	94.7
	tracklet-based	66.1	70.0	62.6	89.5
> 3 s	DS as in [163]	55.8	63.5	49.0	87.4
	tracklet-based	62.5	74.3	52.7	90.1

of 3 seconds, as shown in the middle and bottom sections of Table 3.1.

As expected, the interpolation of the world-fixed 3D location achieves overall better scores than the image-based variant, due to the compensated change in location caused by the movement of the recording vehicle. Still, the tracklet-based generation method surpasses the interpolation results significantly with a **HOTA** value of 66.1 to 45.2 when using all object tracks, and 68.8 to 54.6 when considering only the object tracks which appear in at least two adjoining keyframes. It also outperforms the optimization-based KSP method presented in Section 3.3. Finally, the DS method mimicking the method presented in [163] is clearly outperformed by the tracklet-based method presented in this thesis. As shown in the bottom part of Table 3.1, the tracklet-based method achieves a **HOTA** value of 62.5, outperforming the online baseline by 6.7 points for the **HOTA** metric, and improving for all sub-metrics.

3.6.2. ABLATION STUDIES

In the following, the selected pipeline is analyzed in more detail by decomposing it into its processing stages. At a conceptual level, the full pipeline is referred to as tracklet-based pipeline. At the component level, it consists of a tracklet generation stage (DS), a subsequent tracklet merging stage (**FPC**), and an optional trajectory smoothing. Depending on the processing mode, the method is either applied in a sequence-based manner (S) or processed batch-wise. In the latter case, the results are either aggregated directly using the keyframe information (A) or refined by an additional post-processing smoothing step (P). The superscript K indicates the use of keyframe supervision.

Table 3.2 contains the ablation studies. Several observations can be made. As expected, adding keyframe information (i.e. manually annotated bounding boxes and

Table 3.2: Ablation studies: Effects of keyframing and adding the various stages of the proposed tracklet-based pipeline (tracklet generation, tracklet merging, and trajectory smoothing) on performance. First row *DS* denotes baseline method [187] applied on whole sequence (S) without keyframe information. The other methods are applied batch-wise using keyframe information (i.e. superscript *K*). The latter results are either aggregated “as is” (A) or include post-processing (P) of Subsection 3.5. In batch-wise processing, distinctions are made between tracklet generation (*DS*, Subsection 3.4.1) and tracklet merging (*FPC*, Subsection 3.4.2). All results are reported on the ECP2.0-dense validation dataset.

Method	HOTA↑	AssPr↑	AssRe↑	DetA↑	LocA↑
Sequence-based (S)					
<i>DS</i> [187]	59.5	79.7	65.3	59.7	86.6
Aggregated (A)					
<i>DS</i> ^K	64.2	86.8	70.0	62.5	88.8
<i>DS</i> ^K + <i>FPC</i> ^K	65.1	85.7	73.4	61.8	88.8
Post-Processed (P)					
<i>DS</i> ^K	65.2	87.9	71.0	63.4	89.5
<i>DS</i> ^K + <i>FPC</i> ^K	66.1	86.9	74.5	62.6	89.5

object IDs) - denoted by superscript *K* - improves detection, localization and association performances (columns Association Precision (*AssPr*), Association Recall (*AssRe*), Detection Accuracy (*DetA*) and Localization Accuracy (*LocA*), respectively) and overall *HOTA* score markedly. Furthermore, the proposed three-stage approach with keyframe information (*DS*^K + *FPC*^K, last row) performs best, increasing the overall *HOTA* score by 6.6 points to 66.1 compared to the *DS* method without keyframe information (first row). The beneficial effects of tracklet merging (+*FPC*^K) appear marginal, yet post-processing results in an increase of one *HOTA* point.

Figure 3.4 and Appendix 3.B provide qualitative illustrations of the output generated by the proposed approach (*DS*^K + *FPC*^K with smoothing).

3.6.3. CEILING STUDY (PERFECT DETECTOR)

A ceiling study is conducted to investigate the importance of the detection quality for the intermediate images. Using the validation dataset, the detections are replaced with the actual 2D *GT* bounding boxes, leaving temporal correspondences still to be determined algorithmically. The results for the batch-wise approach with aggregation and without smoothing are shown in Table 3.3 (cf.: middle part in Table 3.2).

When using such positional *GT*, a drastic improvement can be observed in the final performance for all methods. Compared to the results in Table 3.2, the *HOTA* value for the best combination increases from 66.1 to 94.1 (the *DetA* score is slightly below 100, due to possible errors during track initialization and termination). As the *GT* bounding boxes are directly used, the *LocA* score for all methods is 100. However, the *AssRe* value indicates that mistakes are still being made in terms of temporal correspondences. Object ID switches occur mainly when objects are occluded for several consecutive frames, and then they re-appear.

Figure 3.4: Qualitative results of the pseudo GT generation approach proposed in this thesis. A keyframe with the corresponding keyframe annotations is shown on the left and right, while the middle column presents the image in between the keyframes (1 s. apart from keyframes). Track IDs are shown by color-coded boxes. The history of each box in terms of the previous center positions is also shown. Best viewed when magnifying the PDF document. In the top row, a scene from Budapest, with few people during a high yaw-rate turn maneuver: it represents no problem for the proposed approach. The middle row shows a challenging scenario recorded in Karlsruhe, with multiple persons being partly and heavily occluded. Note how the trajectories for the persons are reliably generated also during occlusion (e.g., orange box in the center). Lastly, a challenging scene is shown in the bottom row, recorded at nighttime during rainfall in Zagreb. Here an error in the generated pseudo ground-truth can be observed. An ID-switch happens for the two persons on the left (red box with id 5, brown box with id 17), after being occluded by the crossing pedestrian (blue box id 4).



Table 3.3: Results using ground-truth for all images as detection input (no usage of temporal correspondences). Numbers are shown for batch aggregation (A) without applying smoothing as post-processing. All results are reported on the ECP2.0-dense validation dataset.

Method	HOTA↑	AssPr↑	AssRe↑	DetA↑	LocA↑
<i>DS</i>	52.0	96.9	28.3	96.3	100
<i>DS</i> + <i>FPC</i>	52.4	95.9	29.3	95.0	100
<i>DS</i> ^K	93.7	98.6	89.5	98.5	100
<i>DS</i> ^K + <i>FPC</i> ^K	94.1	97.8	92.0	97.2	100

3.7. DISCUSSION

This chapter introduces a semi-supervised pipeline for creating temporally dense person trajectories from sparse annotations by integrating detection, tracking, and 3D up-lifting. Beyond the individual processing stages, the chapter explored two alternative trajectory generation approaches and design choices, with the overarching goal of reducing annotation effort compared to frame-wise manual labeling while providing the pseudo ground-truth required for large-scale dataset construction.

The experimental investigations in Section 3.6 establish a clear preference for the tracklet-based pipeline presented in Section 3.4 over alternative strategies. When evaluating using the HOTA metric, which jointly captures detection, association, and localization quality, the tracklet-based approach consistently outperforms all other methods as shown in Table 3.1. This includes the graph-based *KSP* method as presented in Section 3.3, which may come as a surprise, given that *KSP* calculates the optimal path solution for a provided cost function. However, it requires careful hand-tuning of cost functions and is less effective in handling of the appearance and disappearance of objects in the scene between keyframes.

On the other hand, the better performance of the tracklet-based method vs. linear interpolation in the ego-motion compensated BEV of the keyframe annotations is to be expected, as the person motion involves non-linearity, which becomes more pronounced at longer time intervals.

The presented tracklet-based method also outperforms the method presented in [163], which uses DeepSORT [187] with an additional post-filtering, for all metrics. Interestingly, when only considering trajectories with a duration of at least three seconds the HOTA value reduces by almost four points.

Despite these improvements over all baselines, the overall performance level highlights remaining limitations of the current pipeline. In particular, detection accuracy remains a dominant bottleneck. This is confirmed by the ceiling study presented in Subsection 3.6.3, where replacing detector outputs with perfect detections leads to a substantial increase in HOTA from 66.1 to 94.1. Also temporal association has its flaws, accounting for the remaining 5.9 HOTA points. The main failure cases involve persons under transient partial or complete occlusions, where no detections are produced in the first place, and/or when temporal correspondence fails during tracklet merging. Also moments of strong vehicle turns tend to produce additional mistakes, as the ego-motion compensation might not perform adequately, leading into an outsized positional resid-

ual between prediction and measurement.

A key strength of the proposed pipeline lies in its modular design. Individual components (including ego-motion compensation, 2D detection, appearance modeling, online tracker, tracklet merging, post-processing) can be replaced independently as improved methods become available.

Finally, it is important to note that the generated labels should be regarded as pseudo ground-truth: although they enable large-scale training and benchmarking, residual errors can propagate from detection noise, occlusion, or calibration uncertainties. The implications of these limitations are discussed in more detail in Chapter 6.

APPENDICES

3.A. APPEARANCE EMBEDDING NETWORK

To leverage the rich visual information to associate persons, an appearance embedding network is employed, based on a person re-identification approach introduced in [59]. For each person measurement, the corresponding image patch is extracted from the bounding box and passed through the network, which maps it into a latent appearance space. A ResNet-50 backbone [55] is used in combination with a multi-task loss comprising the Ranked List Loss [184] and identity classification loss. Following [101] several enhancements are applied, including warm-up learning rate, random erasing augmentation, label smoothing, batch normalization, and a reduced stride in the final convolutional layer. The network hyper-parameter and configuration was chosen, such that the best results are achieved on the ECP2.0-dense dataset (cf. Section 4.6).

The distribution over all appearance distances based on the validation dataset is shown in Figure 3.A.1. Temporal distances are grouped into discrete bins. Using the annotations, distances are further categorized based on whether they correspond to the same or different object. As can be seen, comparing the learned appearance features for the same object in adjacent images results in a significantly reduced distance (i.e., left most distribution). While for larger temporal distances the overlap of the shown distributions increases significantly. This indicates, that the appearance features become less suitable to distinguish different or re-identify the same person over time.

3.B. ADDITIONAL QUALITATIVE RESULTS

An additional qualitative example of the generated pseudo GT is presented in Figure 3.B.1. The Figure displays two keyframes as well as the image between the two keyframes (cf. Figure 3.4). For each image, the generated smoothed 3D positions in the ego-centric BEV are also shown.

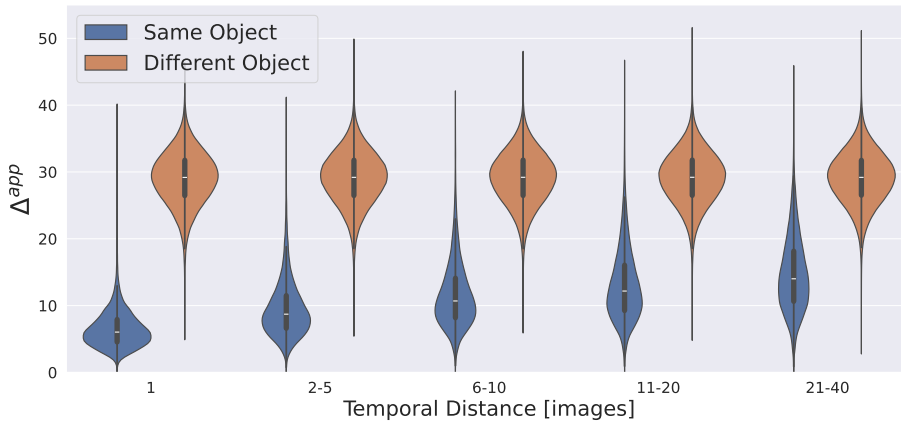


Figure 3.A.1: Distribution of the appearance distances Δ_{app} given two distinct measurements. The distribution of distances for the same object is shown in blue, while distances between non-identical objects are depicted by the orange distribution. Distributions are shown for different temporal distance ranges.



Figure 3.B.1: Additional qualitative results, showing all generated smoothed 3D positions in BEV next to each image. For each object, the trajectory history since the last keyframe is visualized both in the image and BEV. The scene depicts three cyclists at different distances coming toward the ego vehicle, as well as a couple walking on the sidewalk next to the street in front of the car. Best viewed in PDF format with zoom for optimal detail.

4

THE ECP2.0 DATASET FOR PERSON TRACKING IN AD

This chapter introduces the EuroCity Persons 2.0 (ECP2.0) dataset, created using the semi-supervised pipeline presented in Chapter 3. In addition to presenting the dataset and its properties, the quality of the resulting pseudo ground-truth is systematically analyzed. A series of experiments is conducted to assess annotation quality and object-dependent effects. Lastly, the suitability of ECP2.0 as a resource for training and benchmarking modern tracking methods is demonstrated.

4.1. DATA RECORDINGS

The ECP2.0 dataset originates from the same data collection campaign through Europe as the ECP1.0 dataset [13]. The recording vehicle featured a front-facing automotive-grade camera (1920 x 1024) mounted behind the windshield, and a Velodyne HDL-64 LiDAR sensor, mounted on top of the roof. In addition, the vehicle's position, orientation, and velocity were recorded by using an INS sensor. All sensor measurements have been time-synchronized and all sensors have been jointly calibrated (extrinsic calibration) to a common origin. Ego-motion compensation is applied to account for the rolling shutter of the LiDAR sensor and the difference in capture times between the camera and LiDAR. First, in an offline post-processing step the point clouds are rearranged on a measurement slice (vertical) level, such that each point cloud measurement starts at the left edge of the camera's Field of View (FoV). For each image, the rearranged point cloud, with the minimal temporal distance between capture times is selected. Each point in the selected cloud is transformed by the inverse ego-motion (defined by a translation and rotation) of the vehicle between the measurement time of the point and the image. The compensated point cloud is then projected into the camera image using the available calibration information. The final result is shown in Figure 4.1.

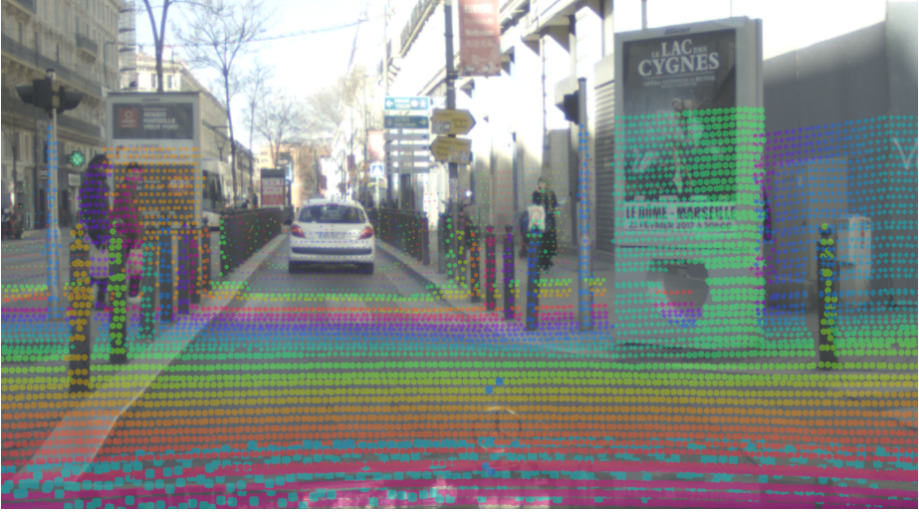


Figure 4.1: Example scene from the ECP2.0 dataset, presenting the compensated point cloud projected into the camera image. Note the good alignment between camera and LiDAR data at the vertical edges of objects.

4.2. ANNOTATIONS

4.2.1. MANUAL ANNOTATIONS

Annotations feature tight bounding boxes capturing the full extent of the object (i.e., if an object is partly occluded the full extent is estimated by the annotator). Only persons with an image height of 40 pixels or larger are annotated. Each person has an assigned class, namely pedestrian or rider. The latter is further categorized based on the vehicle being ridden: bicycle, buggy, motorbike, scooter, tricycle, and wheelchair. During annotation the body orientation of the person is also estimated. Additional annotation information includes a discrete occlusion and truncation level (0% - 10%, 11% - 40%, 41% - 80%, and > 80%, respectively), tags for *standing-or-lying* persons, persons standing *behind-glass*, and *depictions* such as mannequins, posters, or statues. Lastly, areas that contain persons, but for which no individual annotation is possible, are indicated by class-aware ignore regions, referred to as *groups*.

Annotations were performed frame-wise by experienced annotators. Quality standards are implemented by using independent quality control teams. Bounding box annotations need to be accurate within a 2-pixel error bound, while the number of hallucinated and missed objects needs to be below 1%, respectively.

These manual annotations form the keyframes which this thesis uses to generate dense person trajectories. The annotation set is further extended by introducing temporal object correspondences for all keyframe annotations, which are occluded less than 80%. See Appendix 4.A for details on the annotation of temporal object correspondences and the associated GUI tool.

4.2.2. PSEUDO ANNOTATIONS

The pseudo Ground-Truth (GT) for ECP2.0 is generated using the pipeline introduced in Chapter 3. Based on the experimental evaluation in Section 3.6, the tracklet-based trajectory generation approach with post-processing was selected for the final dataset creation. The exact hyperparameters used are provided in Appendix 4.B.

Using this finalized configuration, the following annotations are provided. Each person throughout a sequence is identified by a unique object ID. For the location of each person, the 2D and 3D trajectories are provided. The 2D trajectories include the location, width and height of the person in the image. For the 3D trajectories, the location in the world-fixed Birds-Eye View (BEV) along with the person's height is provided.

Images for which a corresponding 2D measurement is missing for more than one image are labeled as *unsure*, as smoothed trajectories are generated even during occlusions. Similarly, trajectories that do not overlap with at least one keyframe are labeled as *weak*, due to their missing support.

4.2.3. ON THE USE OF 2D BOUNDING BOXES

ECP2.0 provides accurate 2D bounding box annotations to facilitate training and evaluating detection/tracking methods. A conceivable alternative is to annotate 3D bounding boxes in LiDAR data and take the 2D bounding boxes of their projections into the images instead. However, this does not lead to the desired accuracy in practice, due to residual errors in extrinsic calibration and synchronization between the camera and LiDAR sensors, and the fact that the relevant objects' shapes are not really 3D cuboids. For example, consider the following analysis using the WOD dataset [164] (v1.4), which offers both manual 3D and corresponding 2D bounding box annotations for a subset of the objects. Considering all pedestrians and riders with a minimum height of 40 pixel (due to the lack of occlusion information filtering for non-occluded objects was not possible) which are visible in the front-facing camera, results in 467 k pairs of manual 2D and 3D annotations. The comparison between the 2D bounding box resulting from the projected 3D bounding box and the manually annotated 2D bounding box (refer back to Figure 1.4 in Chapter 1 for an illustration) yields an average Intersection-over-Union (IoU) of 50.2. In addition, a mean average error of 15.1 pixels is observed over the bounding box points.

4.3. DATASET STATISTICS

For some basic statistics of the ECP2.0 dataset, refer to the last row of Table 2.2. Figure 4.2 depicts a few relevant object distributions. Figure 4.2a shows the bounding box height distribution. It shows an expected downward trend toward the larger heights; the median and mean bounding box heights are 96 and 128 pixels, respectively. Figure 4.2b shows the trajectory length distribution. Half of the trajectories are shorter than 2.2 seconds (44 images), whereas 25% of the tracks are longer than 4.05 seconds (81 images). Figure 4.2c shows the object density distribution. While most scenes contain a small number of persons, a significant number of scenes feature more than six persons. Specifically, while the mean number of persons per image is 4.4, still 25% of all images contain six or more objects. Figure 4.2d shows the distribution of the object's locations w.r.t. the ego-vehicle in BEV. Given the automotive setting of the recordings, the major-

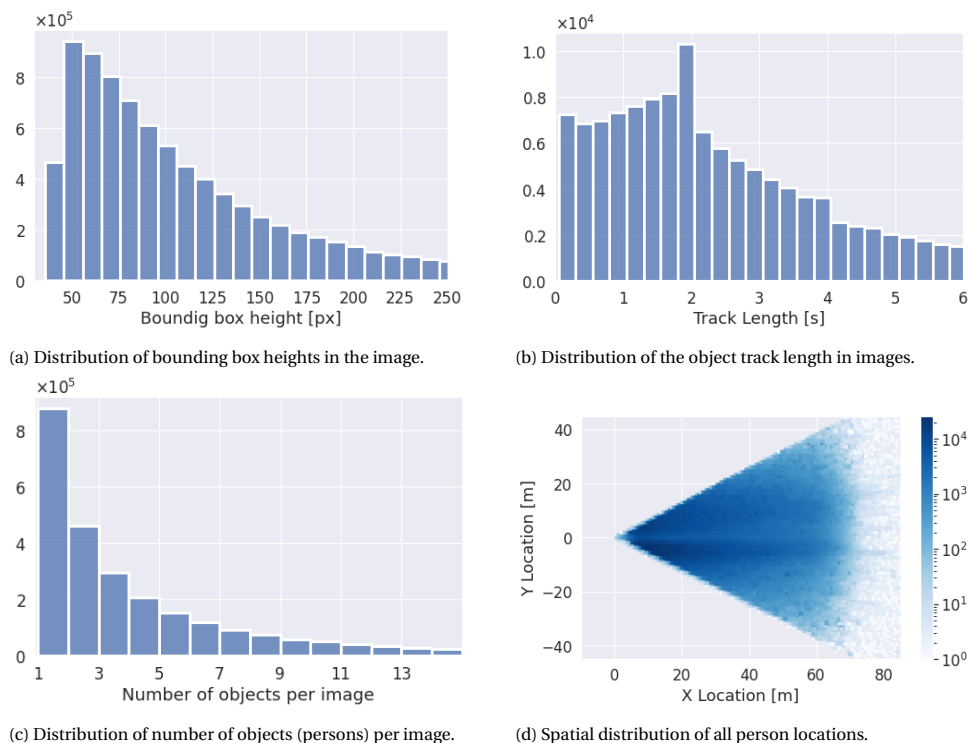


Figure 4.2: Different dataset statistics for all objects contained in the ECP2.0 dataset.

ity of person locations fall into a corridor centered at the origin, which resembles roughly the common street layout. Still, there is a significant amount of data available outside of this corridor, due to intersections, urban areas with wider sidewalks, as well as public squares.

4.4. BENCHMARKING

ECP2.0 follows the same procedure as its precursor ECP1.0 [13] to obtain dataset splits. Each sequence is split into chunks of approximately 20 min., while each chunk is split into a train, validation and test set using a ratio of roughly 60%, 10%, and 30%. All image and ego-motion data is made available. Annotations and generated pseudo ground-truth are only made available for the training and validation sets, not so for the test set. The data is made freely available¹ for academic and non-profit organizations for non-commercial, scientific use. An evaluation server is made available for benchmarking tracking methods, following the Higher Order Tracking Accuracy (HOTA) metric used in this thesis.

4

4.5. TRACKING EVALUATION METRICS

Given the task of linking detections and keyframe annotation over time to generate the desired GT, established tracking metrics are used. Specifically, the recent HOTA metric as presented by Luiten et al. [99]. The HOTA metric evaluates how well trajectories of matching detections align and averages this over all matching detections, while penalizing detections that do not match. In the 2D case, matching is performed using the 2D IoU as similarity score function $\mathcal{S}(c)$ for the candidate pair c containing a prediction and GT. A pair which exceeds the similarity threshold α is considered a true positive (TP). An unmatched GT results in a false negative (FN), while unmatched detections result in a false positive (FP). Using these, the Detection Accuracy (DetA) is calculated by:

$$DetA_{\alpha} = \frac{|TP|}{|TP| + |FN| + |FP|}. \quad (4.1)$$

Similarly, the concept of true positive associations (TPA), false negative associations (FNA), and false positive associations (FPA) is introduced in [99] to calculate the Association Accuracy (AssA) with

$$AssA_{\alpha} = \frac{1}{|TP_{\alpha}|} \sum_{c \in \{TP_{\alpha}\}} \mathcal{A}_{\alpha}(c), \quad (4.2)$$

$$\mathcal{A}(c) = \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|}. \quad (4.3)$$

Using Eq. 4.1 and 4.2 the HOTA score for a single similarity threshold α is given by

$$HOTA_{\alpha} = \sqrt{DetA_{\alpha} \cdot AssA_{\alpha}}, \quad (4.4)$$

¹The ECP2.0 data and an evaluation server for tracking is available from <https://intelligent-vehicles.org/datasets/eurocity/>.

while the final **HOTA** score is calculated by the mean over different thresholds (i.e., $\alpha \in \{0.05, \dots, 0.95\}$) with

$$HOTA = \frac{1}{19} \sum_{\alpha \in \{0.05, \dots, 0.95\}} HOTA_{\alpha}. \quad (4.5)$$

This combined value captures the performance of a method in terms of detection, association, and localization. For a deeper investigation, the separate sub-metrics can be taken into consideration. For instance, the **AssA** can be decomposed into the Association Recall (**AssRe**) and Association Precision (**AssPr**). The association recall decreases when the tracker assigns two different IDs to the same **GT** object, whereas the association precision decreases if the same track ID is used for two different objects. While the ability of a method to precisely localize objects is implicitly captured by the different α thresholds during calculation of the **HOTA** score, Luiten et. al. also introduce the explicit Localization Accuracy (**LocA**) score:

$$LocA = \frac{1}{19} \sum_{\alpha \in \{0.05, \dots, 0.95\}} \left(\frac{1}{|TP_{\alpha}|} \sum_{c \in \{TP_{\alpha}\}} \mathcal{S}(c) \right). \quad (4.6)$$

4.6. VALIDATION DATASET

This thesis uses a subset of ECP2.0, called *ECP2.0-dense*, as validation dataset for pseudo **GT** evaluation. It consists of 227 sequences, with an average duration of 16 s. Manual annotations have been performed at a relatively high frequency of 5 Hz² (compared to the $\frac{1}{2}$ Hz used eventually for the annotation of the entire ECP2.0 dataset), while the annotation protocol was identical to the one used for the manual annotations (see Section 4.2.1) with the only difference being the additionally annotated object IDs. The final annotations for the ECP2.0-dense dataset at 20 Hz are created by linear interpolation in the image; these act as “real” **GT** in the remainder of this chapter, i.e. to evaluate the generated pseudo **GT**. During evaluation, this interpolation is accounted for by ignoring parts of matched tracks which start before the first or end after the last non-interpolated frame of their matched **GT** trajectory, respectively.

The ECP2.0-dense dataset contains a total of approximately 5,500 unique identities, annotated with more than 435,000 (108,750 without interpolation) individual bounding boxes. Quantitative numbers in this chapter (i.e., Table 4.1 - 4.5) are reported using the ECP2.0-dense validation set.

4.7. EVALUATION OF PSEUDO GT

This section presents the experimental evaluation conducted on the ECP2.0 dataset using pseudo **GT** annotations generated with the trajectory generation pipeline selected in Chapter 3. In that chapter, alternative strategies were analyzed and the tracklet-based pipeline was established as the best choice. Accordingly, the experiments in this section focus on analyzing the behavior and properties of the selected pipeline when applied at dataset scale.

²The sampling rate of 5 Hz was chosen according to the Nyquist rate for the human gait cycle, assuming an upper bound of 2 Hz.

Table 4.1: The **HOTA** score of the tracklet-based pipeline, for different keyframe frequencies, on the ECP2.0 dense validation dataset. Next to it, the resulting annotation cost is shown, when extrapolated to the entire ECP2.0 dataset (assuming 2070 k images, see Table 2.2, recorded at 20 Hz, 4.25 annotations per image, and 53 s per annotation (45 s for bounding box and 8 s for object ID)).

Keyframe Freq.	HOTA↑	#Annotations	Annotation Cost (hrs)
-	59.5	0	0
$\frac{1}{4}$ Hz	64.3	110 k	1.6 k
$\frac{1}{2}$ Hz	66.1	220 k	3.2 k
1 Hz	68.3	440 k	6.5 k
2 Hz	71.0	880 k	12.9 k
5 Hz	78.0	2199 k	32.4 k

In particular, Section 4.7.1 examines the effect of the chosen keyframe annotation frequency and its implications for annotation effort and trajectory completeness. Section 4.7.2 analyzes localization accuracy, while Section 4.7.3 examines how pseudo **GT** quality and object statistics vary with object size and keyframe occurrences.

4.7.1. KEYFRAME FREQUENCY

Table 4.1 shows the relation between keyframe frequency (and the number of manual annotations and the annotation cost extrapolated for the entire ECP2.0 dataset, in direct proportion) and the **HOTA** score, as measure of the pseudo **GT** quality using the proposed pipeline. The first table row, denoted by a keyframe frequency of “-”, is the special case of just using the baseline method DS [187]. As can be seen, (near) doubling the keyframe frequency from $\frac{1}{4}$ Hz onwards to 5 Hz, leads to a clear increment in **HOTA** score. A keyframe annotation frequency of $\frac{1}{2}$ Hz (i.e. a 2s time interval) is selected for all subsequent experiments, as a reasonable trade-off between pseudo ground-truth quality (**HOTA**) and annotation cost.

4.7.2. PSEUDO GROUND-TRUTH - LOCALIZATION ACCURACY

A investigation is conducted to quantify the localization accuracy of the generated 2D bounding boxes. A random subset of 450 not occluded pedestrians (two for each sequence in the ECP2.0-dense dataset) was labeled again by three different annotators. Figure 4.3 compares annotator-to-annotator agreement with annotator-to-pseudo-ground-truth agreement. Shown is the average deviation of the median value of the three new annotators for each bounding box side compared to the original annotation from ECP-dense (solid line) and the result of the pipeline proposed in this thesis (dashed line), versus object height.

When comparing the output of the best tracklet-based pipeline and the original ECP2.0-dense annotation, the average deviation of the annotated bounding box sides is 1.6 pixels (standard deviation is 2.2 pixels). The left and right sides of the bounding boxes are less accurate than the top and bottom edges. These findings align with the earlier findings from [13]. A comparison between the median annotation of the three

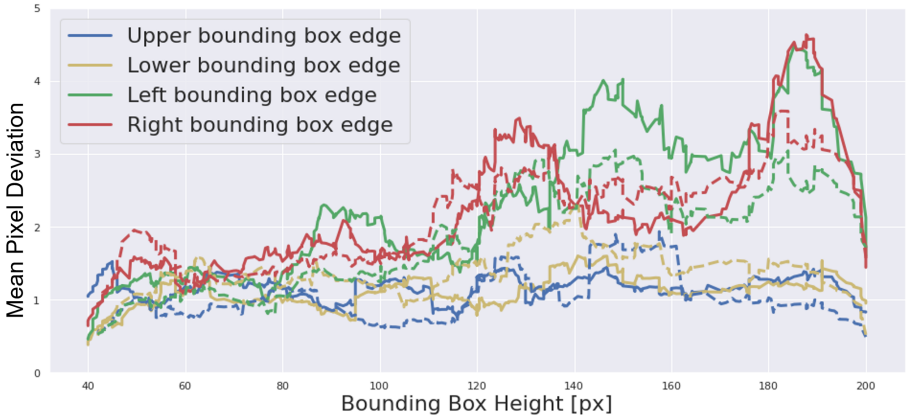


Figure 4.3: Annotator-to-annotator vs. annotator-to-pseudo-ground-truth agreement for bounding box localization. Three additional annotators labeled the bounding boxes of 450 random non-occluded pedestrians. Shown is the deviation for each bounding box side of the annotations from ECP2.0-dense compared to the corresponding median value for the corresponding three additional annotations (solid line) and the result of the proposed approach (dashed line), in dependence of the object height h (averaged over the interval $[h - 20, h + 20]$).

annotators and the original ECP2.0-dense annotations results in an average deviation of 1.7 pixels for the bounding box sides, with a standard deviation of 2.3 pixels. The difference in the annotator-annotator and annotator-to-pseudo-ground-truth deviation is thus not statistically significant. When using the [IoU](#) to compare the median annotator bounding boxes to the ECP2.0-dense annotations, this results in an average [IoU](#) of 87.4 with a standard deviation of 8.7.

4.7.3. [DEPENDENCE ON OBJECT HEIGHT](#)

Table 4.2: Pseudo [GT](#) accuracy ([HOTA](#)) as a function of object size. Listed are also the number of unique persons in the ECP2.0 dataset having a certain set of keyframe annotations. [HOTA](#) value is shown for the tracklet-based pipeline, when reducing the number of objects accordingly on the ECP2.0-dense validation dataset.

Height	#Unique Objects			HOTA↑
# keyframes:	>=0	>=1	>1	
>=40px	228 k	197 k	102 k	66.1
>=60px	177 k	158 k	127 k	68.4
>=80px	140 k	127 k	77 k	70.1
>=100px	113 k	104 k	66 k	71.5

Table 4.2 reports the number of persons for which trajectories are generated, grouped by bounding box height and the number of occurrences in keyframes. By increasing the

Table 4.3: Transfer learning experiments for TrackFormer [110] using the MOT17 and ECP2.0-dense (ECP) datasets.

Trained On	Pre-trained On	Evaluated On (HOTA↓)	
		MOT17	ECP
MOT17	-	57.1	39.4
MOT17	ECP	61.6	42.4
ECP	-	37.6	48.3
ECP	MOT17	40.6	48.1

required minimum bounding box height the overall complexity is reduced, as fewer objects need to be handled and larger boxes in general represent spatial closer measurements and are thus more stable. This also aligns with the HOTA values when reducing the set of objects for trajectory generation and evaluation. If only using and considering objects larger than 100 pixels in height, the HOTA increases by 5.4, with an almost linear progression for heights of 80 and 60 pixels in between. However, this also reduces the number of trajectories by more than 50%.

4.8. GENERALIZATION CAPABILITIES

To assess the generalization capabilities of the dataset, a transfer learning experiment is conducted using TrackFormer [110]. The tracking method is trained separately on MOT17 [112] and on the ECP2.0-dense dataset using the generated pseudo GT (abbreviated as ECP in this section), following its original training procedure [110]. Each trained network is subsequently fine-tuned on the respective other dataset. In the previous sections, the complete ECP2.0-dense dataset is used for validation. Here, it is further split into disjoint train, test, and validation subsets (aligning with the splits from Section 4.4).

The evaluation results on the test split for all four cases are shown in Table 4.3 for both datasets.

As in earlier transfer learning experiments for detection [13], improved tracking performance is observed when pre-training on the target dataset, and best performance when fine-tuning on the target dataset. Specifically, when assessed on MOT17, the model pre-trained on ECP outperforms its counterparts, surpassing the version trained solely on MOT by a margin of 4.5 HOTA points. For evaluation on ECP, the model trained exclusively on ECP yields best performance. Pre-training on MOT17 shows a negligible and even slightly negative impact on the metric (cf. bottom two rows in Table 4.3). The observed increase in performance on MOT17 after pre-training on ECP, as opposed to the lack of improvement on ECP after pre-training on MOT17, highlights the versatility of ECP as a tracking dataset.

4.9. USE OF PSEUDO GROUND-TRUTH FOR BENCHMARKING

Up to this point, the quality of the generated pseudo GT has been evaluated by comparing it to a “real” GT, which was manually annotated at much higher frequency. In this section, the pseudo GT is used to train and evaluate three state-of-the-art tracking

methods. The first method ByteTrack [209], represents a detection-only based tracking-by-detection method. The second method, DeepSORT [187], incorporates additional appearance information during data association. The third method, TrackFormer [110], employs a transformer-based architecture that formulates tracking as an end-to-end learnable task using two adjacent images during training to learn to detect and track objects. All experiments are performed on the ECP2.0-dense dataset to allow for an evaluation using the “real” GT using the splits for training, validation, and testing. For ByteTrack (including its detection network) and TrackFormer, the original hyperparameter configurations from the respective publications are adopted. In the case of DeepSORT, the adapted detection and appearance network as described in Section 3.4.1 is employed.

4

Table 4.4: Performance of tracking methods on the test split of the ECP2.0-dense dataset. *Default* refers to the model provided by the original authors. *Finetuned* refers to the model finetuned using the ECP2.0-dense train split (pseudo GT generated with 0.5 Hz keyframe annotations).

Method	HOTA	DetA	AssA	LocA
Proposed pseudo GT (0.5Hz)	64.6	60.7	69.0	89.2
DeepSORT - Finetuned	58.3	58.3	58.4	86.6
TrackFormer - Default	42.6	43.8	42.5	80.3
TrackFormer - Finetuned	50.9	59.4	44.0	85.5
ByteTrack - Default	55.8	57.5	54.5	82.1
ByteTrack - Finetuned	58.4	60.3	56.6	85.5

Table 4.4 shows the tracking results. As can be seen, using the pseudo GT to finetune increases the HOTA value for TrackFormer and ByteTrack (rows: *Default* vs. *Finetuned*), from 42.6 to 50.9 and 55.8 to 58.4, respectively. Another observation is that after finetuning, the output of the tracking methods remains substantially worse than the proposed pseudo GT, motivating the use of the latter as evaluation “proxy”.

Building on the analysis in Section 4.7.1, tracking performance is now quantified as a function of the keyframe frequency used to generate the pseudo GT for training and evaluation. This investigation is conducted using the TrackFormer method, given its reliance on track information during training (both the spatial bounding boxes as well as the temporal identity information). See Table 4.5. When evaluating with the same “0 Hz” GT quality for training and test split (top row, first column), a clear over-estimation in performance is observed (59.3 vs. 51.6). This highlights the importance of inclusion of manual keyframe annotation to reduce bias in algorithm-based annotation. When evaluating with the “real” GT for the test split (bottom row), a consistent improvement in performance is observed with increasing quality of the training annotations - rising from 49.8 at “0 Hz”, to 50.9 at “ $\frac{1}{2}$ Hz” and 51.6 for “real” GT. Note that this (modest) performance increase is tied to TrackFormer; it does not preclude the possibility that a future, more advanced tracking method might take better advantage of the more accurate pseudo GT provided.

Table 4.5: Relation between GT quality and TrackFormer performance (ECP2.0-dense test split). Column “0 Hz” refers to no use of keyframe annotations (i.e. online use of DeepSORT). The subsequent four columns use pseudo GT at stated keyframe frequencies. Column “GT” involves use of “real” GT (20 Hz). Evaluation either uses the *same* GT quality as used for training (top row) or uses “real” GT (bottom row).

GT Quality (Test split)	GT Quality (Train split)					
	0 Hz	$\frac{1}{4}$ Hz	$\frac{1}{2}$ Hz	2 Hz	5 Hz	GT
Same	59.3	52.5	52.0	50.1	48.1	51.6
GT	49.8	50.9	50.9	51.0	51.3	51.6

4.10. DISCUSSION

This chapter introduces the EuroCity Persons (ECP) 2.0 dataset and its accompanying annotations, which were generated using the semi-supervised pipeline presented in the previous chapter to obtain temporally dense pseudo GT, including 2D bounding boxes and 3D person locations, from sparse manual annotations at keyframes.

Based on the experimental findings on Chapter 3, the tracklet-based pipeline was selected to generate the annotations for ECP2.0. While developed and evaluated in the context of this dataset, the pipeline itself is generic and readily transferable to other datasets with similar annotation characteristics.

The experimental results further showed that interpolation-based annotation result in inferior pseudo GT quality. In fact, linear interpolation between two keyframes is not a viable option in the first place at the chosen annotation frequency of $\frac{1}{2}$ Hz, as only 52% of all persons are present in one keyframe are also present in two keyframes (see Table 4.2). Note that the tracklet-based method presented in this thesis does not have this limitation; it is based on an online tracker where keyframe information provides optional but not necessary information. At the higher keyframe annotation frequency of 2 Hz, similar to [14], linear interpolation becomes more competitive; it obtains a HOTA score of 64.5 vs. 71.0 for the proposed method and less objects are missed by the two keyframe requirement. On the other hand, it does increase annotation cost by a factor of 4.

Despite being generated as pseudo GT, the spatial accuracy of the ECP2.0 annotations is already high, with a LocA value of 89.5. This observation is further supported by the agreement analysis presented in Section 4.7.2, which shows that annotator-to-pseudo-ground-truth agreement is of the same order as annotator-to-annotator agreement. The obtained LocA score (89.5) also aligns well with the IoU / LocA score agreement among annotators of 88 in related large-scale annotation efforts, as reported in [79].

In addition to its scale and diversity, the ECP2.0 dataset demonstrates strong generalization capabilities. As shown in the transfer learning experiment in Section 4.8, models pre-trained on ECP2.0 yield improved tracking performance on external datasets such as MOT17, outperforming models trained solely on the target domain. This highlights the utility of ECP2.0 as a pretraining source for tracking tasks, indicating that the dataset captures a wide range of real-world variations. Moreover, the generated pseudo

ground-truth proved sufficiently accurate for downstream use. In particular, it can serve as supervision for training tracking models, as demonstrated in Section 4.9, where models trained on pseudo-labeled data achieved strong results on the test split. Finally, the pseudo ground-truth was also of sufficient quality for benchmarking purposes. When using pseudo ground-truth without additional keyframe supervision for training and evaluation, the resulting performance was shown to overestimate the true performance when compared to real GT. This bias was found to be significantly reduced when using the proposed pseudo GT, which more closely aligns with real annotations in terms of detection, localization, and association quality. This supports the use of pseudo-labeled data not only for training, but also as a reliable reference for comparing tracking performance across methods.

APPENDICES

4.A. TEMPORAL CORRESPONDENCE ANNOTATION TOOL

As outlined in Section 4.2.1 the annotation set is further extended by introducing temporal object correspondences for all keyframe annotations. Given the already existing bounding box information, the annotation task is defined as follows. Given two adjacent images I_j and I_k and their already established set of bounding boxes, a human annotator is presented a bounding box in I_j and needs to select the corresponding box in I_k depicting the same object - if applicable. This is repeated for all boxes in I_j and for all adjacent image tuples. Missing valid selections (i.e., no corresponding box in I_k for the current person of interest in I_j) are taken into account by introducing a *Non Selectable* option for the labeler. To cope with uncertain situations an *Unsure* selection option is made available to the annotator.

A custom annotation tool was developed to perform the above-mentioned task and is illustrated in Figure 4.A.1. To increase the annotation efficiency appearance-based similarity measures are utilized, based on the embedding network described in Appendix 3.A. Given the current person of interest in I_j up to five similar person boxes are selected from I_k which minimize the distance in the latent appearance feature space. To prevent the introduction of a selection bias toward the top candidate, these suggestions are presented in a randomized order. Thus, the annotator still needs to take the complete scene into account to perform the final selection.

Annotations have been performed by 16 different annotators. An initialization phase was conducted using the same test sequence for each annotator. This enables the calculation of the expected disagreement rates and quantify labeler performance. During the main annotation phase, a repetition factor of 2.0 was applied, i.e., each annotation in every sequence is performed by two different annotators. This results in a disagreement rate of 14.6%. These ambiguous cases are revisited by a third labeler during quality control, which results in a reduced disagreement rate of 6.6%. The final unsure rate is 3.6%. Track IDs are not assigned for unsure and ambiguous annotations. For all remaining objects, the ID given by majority vote was retained.

4.B. HYPER-PARAMETER SETTING

The hyper-parameters used to generate the final ECP2.0 dataset are outlined below.

Batch creation and Detection Input: For the batch creation a keyframe frequency of $f_{KF} = \frac{1}{2}$ Hz is employed, i.e., every 40th image is a keyframe. Batches are created with a length of 41 images, such that each batch starts and ends with one annotated image. A score threshold of 0.8 is applied to the outputs of the retrained person detection network. In addition, all detected persons with a size smaller than 40 pixels in height are ignored. For training the person detection network all available annotations from the keyframes are used.

Tracklet Generation: As described in Section 3.4, two gating thresholds are applied to prevent unlikely matches. The appearance gating threshold is set to $\gamma_{app} = 30$ and spatial threshold to $\gamma_{iou} = 0.7$. For track management, an initialization phase was applied with $n_{init} = 3$. Tracks are deleted after $n_{missed} = 5$ consecutive updates without a corresponding measurement.

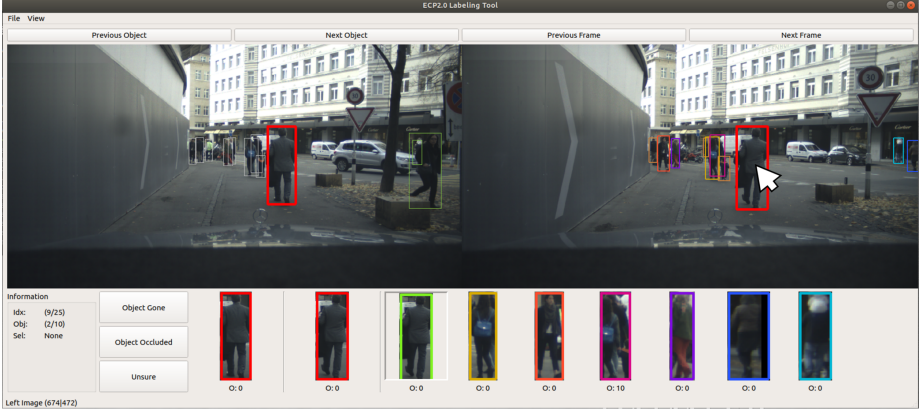


Figure 4.A.1: Annotation tool developed to introduce the temporal correspondence (object IDs) given the already annotated boxes in the keyframes. The tool displays two adjacent images, and presents a person of interest in the left image (red box). The human annotator has to select the corresponding box in the right image (see mouse cursor) - if present. Additionally the tool presents a query of likely candidates by employing image-based appearance information (bottom row and colored boxes).

Table 4.C.1: Density-based evaluation of the proposed approach using the ECP2.0-dense dataset. Each row shows the corresponding evaluation numbers, when discarding images which contain less than the specified number of objects (leftmost column).

Persons/Image	HOTA	DetA	AssA	LocA	#Annotations
≥ 0	66.1	62.6	70.0	89.5	435k
≥ 2	65.9	62.8	69.4	89.5	429k
≥ 5	64.0	62.7	65.6	89.5	374k
≥ 10	59.0	62.1	56.2	89.6	236k
≥ 15	54.6	60.7	49.3	89.5	115k
≥ 20	52.5	58.1	47.5	89.3	53k

Tracklet Merging: The spatial and mean appearance thresholds are set to $\Gamma_{dist} = 10.0$ and $\Gamma_{app} = 50.0$, respectively. Equal weighting is applied for the spatial and appearance distance with $\lambda = 1.0$. In addition, both distance functions are normalized by dividing with their corresponding thresholds, extending Equation 3.8 to:

$$\Delta_{i,j}^{tracklet} = \lambda_1 \Delta_{i,j}^{app} + \lambda_2 \Delta_{i,j}^{dist}. \quad (4.7)$$

With $\lambda_1 = \frac{1}{\Gamma_{app}}$ and $\lambda_2 = \lambda \frac{1}{\Gamma_{dist}}$.

4.C. EVALUATION PER OBJECT DENSITY

To get a deeper understanding, how the density impacts the performance of the presented approach, the following investigation is conducted. The HOTA evaluation as described in Section 4.5 was applied, omitting all images for which the number of objects falls below a pre-defined threshold. The corresponding results are shown in Table 4.C.1.

A noticeable decline in the quality of the generated pseudo GT is observed when only taking images into account with a larger number of objects. As images with a higher number of objects often contain more challenging scenarios. Especially the association accuracy (AssA) drops significantly for higher density values. For crowded scenes, where potentially multiple person boxes overlap, the IoU-based assignment during evaluation becomes more and more ambiguous, leading to an increased number of false matches. In contrast, the DetA remains relatively stable until larger amounts of objects are present. In those cases, smaller sizes and inter-object occlusion leads to a decrease in detection accuracy.

5

LEARNING APPEARANCE EMBEDDINGS FOR PERSON RE-IDENTIFICATION

This chapter introduces a multi-modal method for person re-identification that integrates information from both LiDAR and camera sensors to learn more robust appearance embeddings.

5.1. OVERVIEW

Person Re-Identification ([ReID](#)) methods are commonly used to generate appearance embeddings. Originally developed for surveillance applications with static and overlapping camera views, these methods can also be applied to dynamic, on-board scenarios in intelligent vehicles.

One commonly used approach for multi-object tracking in intelligent vehicles is the tracking-by-detection paradigm [[103](#)], cf. Section [2.2.2](#). In this paradigm, detections are associated over time with the set of existing tracks. A similarity measure is required for associating detections with existing tracks, and for tracking persons, appearance embeddings have proven to be particularly effective [[1](#), [152](#), [176](#), [187](#)].

While image-based [ReID](#) methods are well established, extending these to incorporate multi-modal data from cameras and LiDAR sensors remains underexplored. This is despite the fact that LiDAR sensors are frequently present in intelligent vehicles [[14](#)], and recent research suggests that fusing LiDAR with image data can significantly enhance perception performance [[161](#)].

This chapter presents a novel person [ReID](#) method, which combines image and LiDAR information in an intelligent vehicles setting, as shown in Figure [5.1](#). To effectively combine information from both modalities, several fusion strategies are investigated. While LiDAR data may not provide detailed appearance information, such as clothing or hair color, it can effectively guide visual attention. Additionally, the patch-based na-

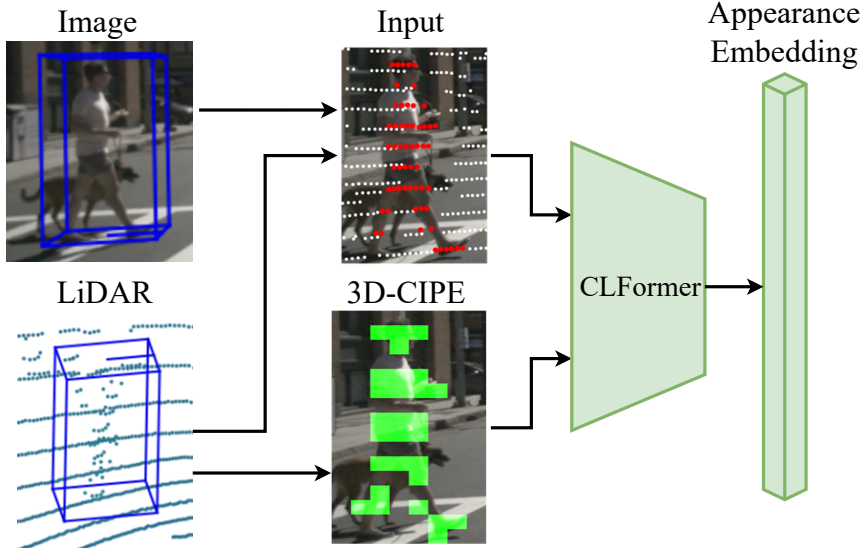


Figure 5.1: Schematic depiction of the proposed novel multi-modal person Re-Identification (ReID) method, using LiDAR point clouds, camera images, and 3D cuboids. Aligned image and LiDAR crops are used as input to generate an appearance embedding for each person. LiDAR points located within the annotated 3D cuboid are used to generate the 3D-CIPE which is added as additional embedding to the transformer.

ture of the transformer architecture facilitates an implicit part-based approach, which is beneficial for handling partial occlusions. To further improve the use of available information, auxiliary tasks are introduced. Lastly, a novel way to integrate object-centered information into the transformer is presented, based on the LiDAR point clouds and available 3D cuboids. The 3D Cuboid-Inclusive Point Embedding (3D-CIPE) incorporates information about whether an image region contains a LiDAR point from within the 3D cuboid to create an abstract object mask which can be embedded into the transformer. The resulting method, referred to as CLFormer, combines the most effective fusion strategy - channel concatenation - with the 3D-CIPE component. To benchmark CLFormer, the nuScenes dataset [14] is repurposed, enabling performance benchmarking and comparison to both baseline and other ReID methods.

5.2. PROPOSED APPROACH

This section describes the proposed appearance embedding network for person ReID (see Figure 5.2 for an overview). The network transforms input sensor data crops into a learned latent appearance space. The distance between two appearance embeddings should be minimal for the same person and maximal for different persons.

5.2.1. CAMERA-ONLY BASELINE

The camera-only ReID baseline uses a transformer-based architecture similar to [57, 87,

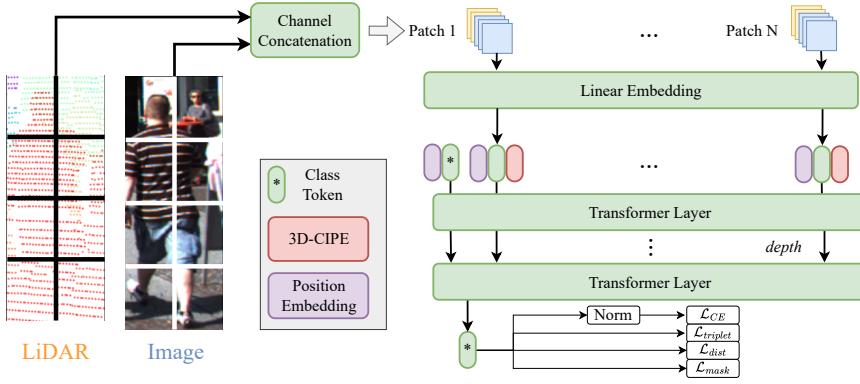


Figure 5.2: The CLFormer method is proposed for multi-modal person re-identification. Each input crop is split into N patches, which are embedded using an initial 2D convolution, augmented by a learnable position embedding (purple box). At this stage, the 3D cuboid-inclusive point embedding is also added (red box). Multi-modal support is added by concatenating the pre-processed LiDAR input to the image crop (see yellow and blue patches at the top). The final class token after the last transformer layer is used as appearance embedding and to compute the losses.

5

134]. Building on the Vision Transformer (ViT) [40], each person crop is resized to a fixed size, resulting in the input $x \in \mathbb{R}^{H \times W \times C}$. The input crop is divided into N patches, each projected into a feature vector of dimension D using a 2D convolution $\mathcal{F}(x)$. To retain spatial information, each patch is augmented with a learnable position embedding and a class token x_{cls} is prepended at the beginning of the feature vector:

$$\mathcal{X}_0 = [x_{cls}, \mathcal{F}(x_1^p), \dots, \mathcal{F}(x_N^p)] + \mathcal{P}, \quad (5.1)$$

with x_i^p being the i -th patch, $i \in [1, N]$, and $\mathcal{P} \in \mathbb{R}^{N+1 \times D}$ (one position embedding for each patch, plus the class embedding). This feature representation is processed by l transformer encoder layers, each consisting of normalization, multi-head self-attention, another normalization, and an MLP layer. The features after the last transformer encoder are normalized and used as input for the ReID tasks.

The ReID task is defined using two losses. First, each unique identity in the training set is treated as an individual class and the network is trained using the cross-entropy loss $\mathcal{L}_{CE}$, encouraging clustering of appearance vectors for identical identities. Second, the triplet loss $\mathcal{L}_{triplet}$ as in [59] is used, to refine the networks ability to distinguish between crops of the same and different classes. The final loss is computed using the weights λ_{CE} and $\lambda_{triplet}$ with:

$$\mathcal{L}_{total} = \lambda_{CE} \cdot \mathcal{L}_{CE} + \lambda_{triplet} \cdot \mathcal{L}_{triplet}. \quad (5.2)$$

5.2.2. ADDING OBJECT POINT INFORMATION AS EMBEDDING

To enhance the model with additional 3D information from LiDAR points and 3D cuboids, the concept of 3D cuboid-inclusive point embedding is introduced. For each LiDAR point, it is determined whether it is contained within the 3D cuboid of the object.

This binary information is compiled for each patch x_i^p processed by the transformer-based model. Patches that contain at least one LiDAR point identified as being inside the object are labeled as *object* patches. The remaining patches are considered as *non-object* patches. A visual representation of this can be seen in Figure 5.1, where LiDAR points located inside the objects 3D cuboid are depicted in red, while the resulting patch-based object mask (i.e., the *object* patches) is shown in green.

The architecture is extended by introducing an additional object mask embedding. A learnable parameter is added to each patch to differentiate between *object* and *non-object* patches. Specifically, Equation 5.1 is extended as follows:

$$\mathcal{X}_0 = [x_{cls}, \mathcal{F}(x_1^p) + \omega_1, \dots, \mathcal{F}(x_N^p) + \omega_N] + \mathcal{P}, \quad (5.3)$$

with

$$\omega_i = \begin{cases} \omega^O, & \text{if } x_i^p \text{ is object patch,} \\ \omega^B, & \text{if } x_i^p \text{ is non-object patch.} \end{cases} \quad (5.4)$$

Here, ω^O and ω^B are the learnable parameters ($\mathbb{R}^D$) for object and non-object patches, respectively. The augmented feature representation $\mathcal{X}_0$ is then fed into the subsequent transformer layers.

5.2.3. EXTENSION TO MULTI-MODAL INPUT

The previously described baseline, which uses only image data as input, is extended to incorporate multi-modal sensor data, specifically camera and LiDAR. In the following, the necessary pre-processing is described and three distinct fusion strategies are proposed.

First, the sensor information is aligned into a common space by projecting the LiDAR points into the camera crop. For the camera image, the temporally closest available point cloud is selected. Using extrinsic calibration information, the LiDAR points are projected into the corresponding camera axis. Each point is then projected into the image space using the intrinsic calibration of the camera. Thus, for each LiDAR point, its position in the camera image (u, v) , its original 3D position (X, Y, Z) converted into the camera axis, and its intensity I are obtained. Points whose image positions fall outside the image crop are discarded. The Z value for each point is referred to as its *depth* value in the remainder of this chapter. Figure 5.3 provides a visual representation of this procedure. Given the sparsity of the data, interpolating the LiDAR data is proposed, as shown in Figure 5.3e.

This chapter proposed three distinct fusion strategies to integrate this multi-modal data into the network as shown in Figure 5.4. The first strategy, channel concatenation, involves concatenating the LiDAR data as additional channels to the input image crop (cf. variable "C" in Sec. 5.2.1), requiring the first convolution layer to process both modalities jointly. The second strategy uses separate embedding layers for each modality, splitting $\mathcal{F}$ into $\mathcal{F}_{im}$ for the image and $\mathcal{F}_L$ for the LiDAR data. The resulting feature vectors can be fused by summation, multiplication, a fully connected layer, or concatenation. For both fusion strategies, the resulting embedded vector $\mathcal{X}_0$ is used as input for the transformer encoders, aligning with the camera-only architecture. The third strategy, cross-modal attention fusion, employs separate transformer backbones to embed

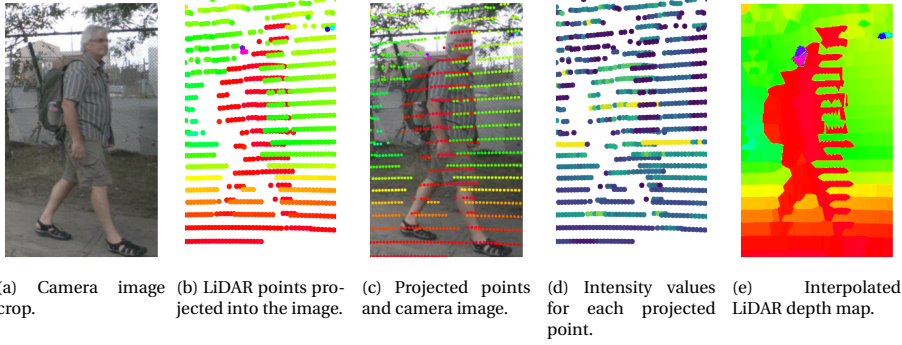


Figure 5.3: Example camera and LiDAR crops. The distance value of the points is shown with red (close) to green (distant).

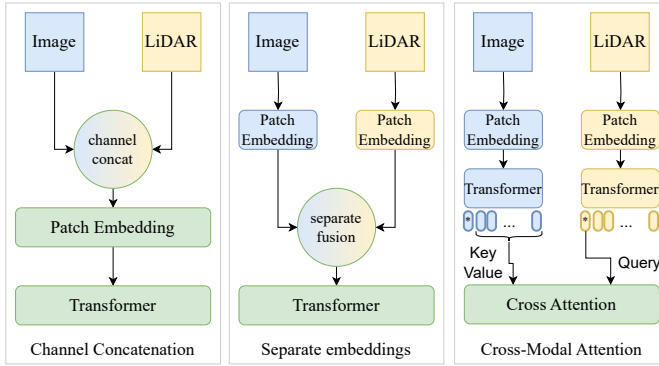


Figure 5.4: This chapter proposes and investigates three different multi-modal fusion approaches to combine LiDAR and image data.

the camera and LiDAR data independently. A cross-attention head is used to exchange information between the final camera and LiDAR tokens. Either the LiDAR or camera tokens are used as the Query, while the remaining modality tokens serve as Key and Value. For the Query, only the class token is used, while for the Key-Value modality, experiments are performed using the patch tokens alone and in combination with the class token.

5.2.4. AUXILIARY TASKS

Auxiliary tasks are also explored to leverage the spatial information from the LiDAR modality. First, a regression task is added to estimate the mean depth of each person crop, using the annotated 3D box as the target depth. This task is trained with the mean squared error loss $\mathcal{L}_{dist}$. Second, a patch-wise depth delta relative to the target depth is estimated. For each patch, the mean LiDAR point depth is calculated and the delta is determined as the difference between this mean depth and the target depth. The L1 loss $\mathcal{L}_{mask}$ is used to compare the networks per-patch delta estimation with the target

Table 5.1: Statistics of the dataset used for the experiments. The nuScenes-ReID dataset is introduced, created from the nuScenes driving [14] dataset. Each unique person identity is given by their annotated track ID.

Dataset	Split	Identities	Crops	Cameras
nuScenes-ReID	train	5,284	53,794	6
	query	1,029	2,421	6
	gallery	1,029	8,220	6

values. The final loss function is extended to include these auxiliary tasks using weights λ_{dist} and λ_{mask} :

$$\begin{aligned}\mathcal{L}_{total} = & \lambda_{CE} \cdot \mathcal{L}_{CE} + \lambda_{triplet} \cdot \mathcal{L}_{triplet} \\ & + \lambda_{dist} \cdot \mathcal{L}_{dist} + \lambda_{mask} \cdot \mathcal{L}_{mask}.\end{aligned}\quad (5.5)$$

5

5.3. EXPERIMENTS

5.3.1. DATASET

The experiments focus on using the nuScenes driving dataset [14], given its size, quality, and the availability of image and LiDAR data. The dataset offers benchmark protocols for the tasks of object detection, tracking, and prediction among others, but no benchmark setting is provided for person re-identification. To create a ReID dataset version, all annotated persons in the dataset are used. For each person (i.e., pedestrians and riders) their corresponding unique id (given by their track id), image and LiDAR crop are extracted. The train/val split provided by the nuScenes dataset is followed to separate train, gallery and query splits. For the latter two, the validation split is subdivided.

The resulting *nuScenes-ReID* dataset contains the images from all cameras, and was used as the dataset for all experiments reported in the following sections. To improve dataset quality, only crops are considered which are larger than 40 pixels in height, 20 pixels in width, contain more than 25 LiDAR points, and are annotated with an object visibility of 40% or more. The final number of identities and crops is reported in Table 5.1 for the dataset.

5.3.2. EVALUATION METRICS

The ReID model is evaluated using the commonly used mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) [213]. Evaluation is done using the query and gallery subsets. The appearance embeddings for all samples from the gallery and query are generated. Each query embedding is compared to all gallery embeddings.

The CMC metric specifies the percentage in which the same identity as the query is contained in the closest k selected gallery samples. Commonly, the *rank-1* is used in the ReID setting. To calculate the mAP, the average precision (AP) for each sample is used. The AP provides a measure of how well the network is able to identify the same person, taking the recall and precision into account.

Table 5.2: Experimental results of various ReID methods on the NuScenes-ReID dataset. The first section presents different configurations of the image-based transformer baseline. The middle section displays the results of CNN-based approaches, while the final section shows transformer-based approaches. Each metric is reported as the mean and $\pm std$ from four different training runs.

Model	mAP	Rank-1
Baseline - ViT	77.56 $\pm$ 0.06	91.14 $\pm$ 0.02
Baseline - DeiT	77.10 $\pm$ 0.11	91.10 $\pm$ 0.10
Baseline - CrossViT	73.91 $\pm$ 0.13	89.46 $\pm$ 0.06
ViT (patchsize 14)	78.26 $\pm$ 0.04	91.22 $\pm$ 0.09
ViT (patchsize 12)	78.52 $\pm$ 0.13	91.16 $\pm$ 0.29
BoT [101]	75.59 $\pm$ 0.16	91.29 $\pm$ 0.30
SBS [56]	76.38 $\pm$ 0.02	91.26 $\pm$ 0.14
AGW [198]	76.48 $\pm$ 0.16	91.70 $\pm$ 0.15
MGN [175]	77.23 $\pm$ 0.12	91.52 $\pm$ 0.21
LoGoViT w/o PM [134]	77.20 $\pm$ 0.05	90.52 $\pm$ 0.12
TransReID [57]	78.80 $\pm$ 0.07	91.43 $\pm$ 0.30
DC-Former [87]	78.89 $\pm$ 0.09	91.59 $\pm$ 0.23
CLFormer (proposed)	79.85 $\pm$ 0.10	92.76 $\pm$ 0.12

5.3.3. EXPERIMENTAL SETUP

For all experiments $H = 256$, $W = 128$, $D = 768$, and $l = 12$ are used. If not specified otherwise a quadratic patch size of 16 pixels is used, with a stride of 16. During training random input augmentations are employed for both LiDAR and image data, including horizontal flipping, padding, and cropping. A batch size of 256 is used. The optimizer is SGD with a momentum of 0.9, weight decay of $1e^{-4}$, and initial learning rate of 0.032. A cosine learning rate decay is used, with a warm-up rate of 40 epochs, and is performed for a total of 120 epochs. Implementation is based on PyTorch. Each experiment is run using a single NVidia RTX A6000 GPU. The weights loaded for the ViT backbone are pre-trained on ImageNet-21K and finetuned on ImageNet-1K. For the multi-modal input, the weights for the initial convolution layer ($\mathcal{F}$) are loaded for the image channels - if applicable. LiDAR channel weights are initialized randomly.

To ensure the reliability of experimental results, each training is conducted four times, each with a unique random seed. The seed initializes the state of all random number generators in the training process. This procedure enables computation and reporting of the mean and standard deviation for all evaluation metrics. Early stopping based on the validation mAP value was used to prevent overfitting.

5.3.4. COMPARISON TO BASELINES AND STATE-OF-THE-ART

The proposed method CLFormer is compared to various baselines and state-of-the-art methods in Table 5.2. In the first section of the table, results are presented for different configurations of the image-based transformer baseline. Experiments were conducted with different backbones (ViT [40], DeiT [169], and CrossViT [23]) and varying patch

sizes. Consistent with expectations and the findings in [57], the best performance is observed with ViT using the smallest patch size and a stride of 12 pixels, increasing the **mAP** from 77.56 to 78.52 points. This demonstrates the effectiveness of using finer granularity in patch sizes for enhancing the models discriminative power - at the cost of higher computational cost. Additionally, the DeiT model, although slightly behind ViT, shows competitive performance with an **mAP** of 77.10.

The center of the table shows results for CNN-based methods. For all four methods - BoT [101], SBS [56], AGW [198], and MGN [175] - a pretrained ResNet-50 backbone network is used along with Instance-Batch Normalization (IBN) [129]. As expected, all four methods lag behind the transformer-based baseline (first section), given their less efficient backbone architecture. Among these methods, MGN achieves the best performance with its global and local branches, yielding an **mAP** of 77.23 and a Rank-1 of 91.52.

In the last section, results are presented for more recent transformer-based networks. The LoGoViT network [134] without their Patch Modification (PM) module falls behind the transformer baseline by 0.36 **mAP** points, achieving an **mAP** of 77.20 and a Rank-1 accuracy of 90.52. Both TransReID [57] and DC-Former [87] demonstrate improved **ReID** performance, with increases of 1.24 and 1.33 **mAP** points, respectively. TransReID achieves an **mAP** of 78.80 and a Rank-1 accuracy of 91.43, while DC-Former achieves the second-best **mAP** of 78.89 and a Rank-1 accuracy of 91.59.

The last row shows the proposed CLFormer method, which combines multi-modal input concatenation, the proposed 3D-CIPE, and additional auxiliary losses. CLFormer achieves the best results for both **mAP** and Rank-1 metrics, with an **mAP** of 79.85 and a Rank-1 accuracy of 92.76. This represents an increase of 0.96 in the **mAP** over the second-best method (DC-Former) and 2.29 over the baseline (ViT).

5.3.5. IMPACT OF DIFFERENT MULTI-MODAL FUSION STRATEGIES

The results in Table 5.3 indicate that the fusion strategy significantly impacts the performance of the **ReID** method. Among the separate embedding strategies, summation fusion (*S-Sum*) achieved the highest **mAP** of 77.81% and a Rank-1 accuracy of 91.14%. This performance, along with the observed drop in the other separate fusion settings, is attributed to the misalignment between the hidden dimensions of the transformer and the pre-trained model weights optimized for image-only data. Additionally, a nearly two orders of magnitude higher standard deviation is observed in these settings, indicating potential instabilities during training. For CrossAttention strategies, the configuration with LiDAR to Camera attention and class tokens ($CA^{L \rightarrow C} + \text{cls}$) performed well, achieving an **mAP** of 75.56% and a Rank-1 accuracy of 90.72%. This result is consistent with the expectation that using the LiDAR token as the query allows the extraction of meaningful features from the image tokens used as keys and values, compared to the reverse configuration, which indicates a lack of distinct appearance information in the LiDAR data. Including the class token in the keys and values has no statistically significant impact. Interestingly, the concatenation fusion strategy (*Concat*) with interpolated input and both Depth and Intensity channels (*D+I*) achieved the highest overall performance, with an **mAP** of 77.9% and a Rank-1 accuracy of 91.14%. This suggests that simple concatenation of multi-modal inputs can be highly effective when combined with interpolation. Consequently, *Concat* is used as the fusion strategy for subsequent experiments.

Table 5.3: Results for various multi-modal fusion configurations using different LiDAR channels (*D*: Depth, *I*: Intensity), input types (*sparse* or interpolated (*Interp.*)), and fusion strategies. Fusion strategies include input channel concatenation (*Concat*), separate embedding (*S*) with linear, product, summation, or concatenation fusion, and CrossAttention (*CA*) with Camera to LiDAR (*C→L*) or LiDAR to Camera (*L→C*) attention, with or without class tokens (*cls*). See Section 5.2.3. Each metric is presented as mean $\pm$ std.

Input	Ch.	Fusion	mAP	Rank-1
Interp.	D+I	S-Product	65.46 $\pm$ 3.68	86.61 $\pm$ 1.37
Interp.	D+I	S-Linear	63.78 $\pm$ 2.51	85.88 $\pm$ 1.42
Interp.	D+I	S-Sum	77.81 $\pm$ 0.07	91.14 $\pm$ 0.25
Interp.	D+I	S-Concat	59.68 $\pm$ 0.35	84.24 $\pm$ 0.47
Interp.	D+I	CA ^{C→L} + cls	30.87 $\pm$ 6.12	57.79 $\pm$ 7.00
Interp.	D+I	CA ^{L→C} + cls	75.56 $\pm$ 0.17	90.72 $\pm$ 0.32
Interp.	D+I	CA ^{L→C}	75.53 $\pm$ 0.07	90.73 $\pm$ 0.15
Sparse	D+I	Concat	77.59 $\pm$ 0.03	90.88 $\pm$ 0.27
Interp.	I	Concat	77.69 $\pm$ 0.09	90.89 $\pm$ 0.15
Interp.	D	Concat	77.66 $\pm$ 0.09	90.96 $\pm$ 0.20
Interp.	D+I	Concat	77.90 $\pm$ 0.05	91.14 $\pm$ 0.18

5.3.6. ABLATION STUDY: AUXILIARY TASKS AND 3D-CIPE

The integration of depth-related auxiliary tasks into the proposed network was based on the hypothesis that such tasks would steer the network toward effectively utilizing the LiDAR data in a multi-modal context. To evaluate this, an ablation study was conducted using both the image-only baseline and the multi-modal configuration, which incorporates input concatenation with interpolation. These experiments were performed with and without the auxiliary tasks. Furthermore, the impact of introducing the proposed 3D Cuboid-Inclusive Point Embedding (3D-CIPE) was investigated. The outcomes of these experiments are summarized in Table 5.4.

The upper and central sections of the table present the incremental activation of the additional distance and patch-delta estimation tasks for the image-only and multi-modal models, respectively. Contrary to initial expectations, the incorporation of these tasks did not enhance the [ReID](#) performance; rather, a marginal decline was observed. This trend was consistent across both the image-only and multi-modal models. Specifically, the inclusion of the distance estimation task did not yield a notable improvement, while the patch delta estimation task seems to contradict the training slightly, leading to decreased performance.

In contrast, the implementation of the proposed 3D-CIPE results in a significant improvement in [mAP](#). The image-only baseline experiences an increase in [mAP](#) from 77.6 to 79.5 with the integration of the LiDAR-based embedding. A similar improvement is observed for the multi-modal model, with an [mAP](#) rise of nearly 2 points.

Table 5.4: Results of the ablation study for additional auxiliary tasks and 3D-CIPE. The performance is shown for the image-only baseline, and the proposed multi-modal model using both LiDAR channels interpolated. The columns *D*, *P*, and *E*, indicate if the distance estimation, patch-delta estimation, and 3D-CIPE was added, respectively. For each metric the mean and $\pm std$ from four different runs is shown.

Model	D P E	mAP	Rank-1
Image	- - -	77.56 $\pm$ 0.06	91.14 $\pm$ 0.02
Image	✓ - -	77.54 $\pm$ 0.08	90.88 $\pm$ 0.24
Image	- ✓ -	77.45 $\pm$ 0.06	90.95 $\pm$ 0.23
Image	✓ ✓ -	77.47 $\pm$ 0.04	90.90 $\pm$ 0.19
Multi-Modal	- - -	77.90 $\pm$ 0.05	91.14 $\pm$ 0.18
Multi-Modal	✓ - -	77.87 $\pm$ 0.08	91.32 $\pm$ 0.16
Multi-Modal	- ✓ -	77.66 $\pm$ 0.05	91.15 $\pm$ 0.10
Multi-Modal	✓ ✓ -	77.73 $\pm$ 0.07	91.12 $\pm$ 0.23
Image	- - ✓	79.54 $\pm$ 0.07	92.78 $\pm$ 0.12
Image	✓ ✓ ✓	79.45 $\pm$ 0.05	92.48 $\pm$ 0.30
Multi-Modal	- - ✓	79.78 $\pm$ 0.11	92.77 $\pm$ 0.25
Multi-Modal	✓ ✓ ✓	79.85 $\pm$ 0.10	92.76 $\pm$ 0.12

5.4. DISCUSSION

The experimental results indicate that the proposed CLFormer method offers clear performance gains over both the baseline and recent state-of-the-art approaches. Increases in mAP and Rank-1 accuracy highlight the enhanced discriminative power of the learned appearance embeddings. Notably, the relatively low standard deviations across multiple training runs suggest that the method is both stable and reproducible, reinforcing the validity of the observed improvements (cf. Table 5.2).

A critical aspect of the proposed method is the incorporation of multi-modal fusion strategies. Among the strategies investigated, simple input channel concatenation with interpolated LiDAR data proved most effective (cf. Table 5.3). This observation suggests that, in this setting, a straightforward fusion of modalities is sufficient to capture complementary information, thereby improving re-identification performance. In contrast, more complex fusion strategies such as separate embedding with various fusion operations or cross-modal attention did not consistently yield additional benefits. This may be attributed to the intrinsic misalignment between the pretrained image features and the additional LiDAR information, which may reflect limitations in the scale of the experimental dataset or indicate that the current architectural choices are not yet optimally aligned to fully leverage complementary modality cues.

The introduction of the 3D Cuboid-Inclusive Point Embedding (3D-CIPE) component further improved the performance of the multi-modal approach. By encoding binary object masks derived from LiDAR points and 3D cuboids, the network benefits from an enriched spatial representation that guides the attention mechanism in the transformer. This enhancement is evidenced by the noticeable increase in mAP upon integrating the 3D-CIPE (cf. Table 5.4), despite its reliance on a relatively simple binary sig-

nal. However, the use of a binary object mask also raises questions regarding potential limitations, such as the loss of more nuanced geometric information, which could be explored in future work.

Conversely, the auxiliary tasks designed to estimate depth and patch-wise delta values did not yield improvements and, in some configurations, even slightly degraded the re-identification performance (cf. Table 5.4). This outcome may indicate that the additional loss terms compete with the primary objective or that the current formulation of these auxiliary tasks is not optimally aligned with the models capacity to exploit LiDAR data.

When compared to conventional CNN-based methods, transformer-based architectures such as CLFormer demonstrate several advantages, including better adaptability to variable patch sizes and improved handling of occlusions. The experimental comparisons also underscore that, while CNNs remain competitive, the transformer paradigm appears more naturally suited to the integration of multi-modal inputs. Notably, Table 5.2 clearly shows that all transformer-based configurations outperform the CNN-based baselines in both mAP and Rank-1 accuracy.

6

CONCLUSION AND FUTURE WORK

This thesis addressed the problem of vision-based multi-object tracking and person re-identification for automated driving in urban environments. To this end, a semi-supervised pipeline for generating temporally dense annotations was developed, enabling the creation of large-scale datasets without exhaustive manual labeling. Building on this, the EuroCity Persons 2.0 dataset was introduced, offering diverse pseudo ground-truth annotations for benchmarking person tracking in traffic. Furthermore, a novel multi-modal appearance model was proposed, combining camera and LiDAR data in a transformer-based framework to improve person re-identification under challenging conditions. Together, these contributions advance the robustness of vulnerable road user perception and provide both methodological insights and data resources to stimulate future research in intelligent vehicles.

6.1. CONCLUSION

This section draws conclusions for the methodical chapters presented in this thesis, based on the Research Questions (RQs) formulated in Section 1.1.5. After discussing the chapter-wise conclusions the section concludes by reflecting on how this thesis addressed the individual research questions.

SEMI-SUPERVISED CREATION OF AN AD DATASET

Chapter 3 introduced a semi-supervised pipeline for dataset creation tailored to person tracking in automated driving, directly addressing RQ2 by investigating how temporally dense trajectory supervision can be generated. Sparse manual 2D annotations at keyframes were complemented with automated methods to produce temporally dense person trajectories, substantially reducing annotation costs.

The 2D image annotations were enriched with LiDAR input via a novel 3D uplifting procedure, which leveraged 2D boxes, prior knowledge about human shape, and a constrained clustering algorithm to estimate 3D positions. Ego-motion information was applied to transform these positions into a consistent world-fixed coordinate system.

For batch-wise processing, sequences were split into segments delimited by keyframes. Two approaches were presented for track generation within each batch: a graph-based optimization using the k-shortest path algorithm (cf. Section 3.3), and a tracklet-based method that combined tracklet generation with merging (cf. Section 3.4). After per-batch track generation, a final aggregation step and model-based post-processing were applied to obtain the final person trajectories.

Comparing the performance of the generated pseudo ground-truth obtained with the two different track generation methods showed that the tracklet-based approach clearly outperformed the graph-based formulation. The best pseudo ground-truth achieved a Higher Order Tracking Accuracy (HOTA) track quality score of 66.1. Incorporating information from keyframe annotations into tracklet generation, tracklet merging, and track smoothing was shown to be critical, as demonstrated by the ablation studies. Significant improvements over 3D interpolation, graph-based optimization and online tracking baselines were observed, exceeding five HOTA points.

ECP2.0 DATASET

Chapter 4 introduced the new EuroCity Persons (ECP) 2.0 dataset, the largest automotive image dataset for persons in traffic by nearly an order of magnitude, which comes with remarkable data diversity. The novel semi-supervised pipeline presented in Chapter 3 was used to generate the temporally dense pseudo ground-truth (i.e. 2D bounding boxes, 3D person locations) from sparse, manual annotations at keyframes.

Transfer learning experiments highlighted the generalization potential of the dataset, and the benefits of the generated pseudo ground-truth for training and testing of state-of-the-art tracking methods were demonstrated. In this context, Chapter 4 addresses RQ3 by showing that the generated annotations are sufficiently reliable to support both training and benchmarking of modern tracking methods.

The scope of this thesis was to obtain accurate 2D bounding box annotations for large-scale image datasets in order to enable reliable detection and tracking evaluation. As detailed in Section 4.2.3, relying on projected 3D annotations for image-domain supervision leads to considerable 2D localization errors. The analysis on WOD reports an average Intersection-over-Union (IoU) of 50.2 between projected 3D boxes and manual 2D annotations, with a mean corner-point error of 15.1 pixels. In contrast, the pseudo ground-truth generation employed for ECP2.0 operates directly on image-based annotations and achieves a mean deviation of 1.6 pixels, as reported in Subsection 4.7.2, indicating substantially increased 2D localization accuracy.

The semi-supervised pseudo ground-truth generation approach with a keyframe interval of 2 s resulted in a reduction in annotation cost of about factor 34 (instead of annotating 40 frames at 45 s, one keyframe at 45 s is annotated and one set of object ID correspondences between keyframes at 8 s, on average). This is useful, as it facilitated the current ECP2.0, which by its sheer size (an order of magnitude larger than other existing public datasets) and diversity, can hopefully stimulate research in person tracking and motion prediction. Yet a keyframing approach will principally reach its limits at intervals not much larger than the one chosen, just by the fact that object tracks will increasingly fall “in between” keyframes.

CLFORMER: MULTI-MODAL PERSON RE-IDENTIFICATION

Chapter 5 presented CLFormer, a novel transformer-based method for multi-modal person Re-Identification (ReID) using camera and LiDAR data. By combining camera and LiDAR data, the method addressed a gap in ReID research, which has traditionally focused on image-only approaches. The nuScenes dataset was repurposed for the ReID task, enabling a thorough evaluation of the proposed method. Various fusion strategies were explored, with the early concatenation of additional LiDAR channels found to be the best-performing approach.

Experiments demonstrated that the integration of LiDAR data significantly improves ReID performance. The proposed 3D-CIPE module, which incorporates object-centered information from LiDAR point clouds and 3D cuboids into the transformer architecture, yielded substantial gains in mean Average Precision (mAP) and Rank-1 accuracy. Conversely, the inclusion of depth-related auxiliary tasks, intended to guide the network to focus more on the LiDAR input, did not result in increased ReID performance. These findings provide valuable insights into the complexities of multi-modal data integration and suggest areas for further optimization.

Overall, CLFormer highlights the potential of leveraging complementary sensor modalities for person ReID in intelligent vehicle contexts. In this way, Chapter 5 addresses RQ4 by demonstrating that combined camera and LiDAR information can lead to measurable improvements. At the same time, the observed reliance on image-only pre-trained backbones and the limited benefit of certain fusion strategies point toward remaining challenges and opportunities for future research in multi-modal ReID.

REVISITING THE RESEARCH QUESTIONS

The research questions formulated in Section 1.1.5 were introduced to structure the contributions of this thesis and connect the methodological advances to the challenges of visual tracking in urban automated driving. For completeness and to facilitate readability, the individual RQs are repeated here:

- **RQ1:** What methods are used for visual tracking and person re-identification of Vulnerable Road Users (VRUs), and what limitations emerge when these methods are considered in the context of urban automated driving?
- **RQ2:** How can temporally dense trajectory supervision be generated from sparse 2D annotations, using only on-board sensors?
- **RQ3:** How reliable is the generated pseudo ground-truth for training and benchmarking modern tracking methods?
- **RQ4:** Can person re-identification be improved by combining information from camera and LiDAR?

The first research question RQ1 examined the limitations of existing visual tracking and person ReID approaches. While this question cannot be fully answered as long as visual object tracking itself remains an open research question, this thesis provides a partial answer by clarifying where current approaches face fundamental constraints. The

literature review in Chapter 2 shows that recent advances are strongly driven by increasingly large annotated datasets, indicating that data availability has become a key bottleneck.

RQ2 and RQ3 directly address this bottleneck by asking whether manual annotation effort can be substantially reduced while still enabling reliable training and evaluation of tracking methods. Chapter 3 demonstrated that the proposed semi-supervised pipeline can generate high-quality pseudo Ground-Truth (GT) from sparse keyframes at a fraction of the manual effort, while preserving localization and tracking accuracy. Chapter 4 extends this contribution through the introduction of the ECP2.0 dataset, which operationalizes this pipeline at scale and provides a diverse benchmark for person tracking in urban traffic. Beyond answering these research questions, ECP2.0 constitutes a major contribution of this thesis by enabling systematic evaluation of tracking methods in settings that were previously underrepresented.

RQ4 investigated whether person ReID performance can be improved through multi-modal fusion of camera and LiDAR data. The results in Chapter 5 confirm that adding geometric cues from LiDAR leads to measurable and consistent gains over image-only baselines. At the same time, the observed saturation of more complex fusion strategies highlights current limitations in aligning heterogeneous representations and leveraging pre-trained image backbones for truly balanced multi-modal reasoning.

Taken together, these partial answers contribute to answering the overarching RQ of how visual tracking of VRUs in urban traffic can be improved under practical constraints. This thesis shows that such improvements are achievable in practice by reducing the dependency on fully manual annotations and by strengthening identity reasoning through complementary sensor cues. In particular, the proposed semi-supervised annotation pipeline and the resulting ECP2.0 dataset directly tackle the challenge of high annotation effort and limited dataset scale, while the multi-modal ReID experiments demonstrate that sensor fusion can partially compensate for appearance ambiguities.

At the same time, these improvements do not resolve all challenges identified in Section 1.1.4. Persistent identity fragmentation under long occlusion, extreme crowding, and adverse conditions such as nighttime operation remain largely unaddressed. Moreover, the remaining performance gap to idealized manual GT indicates that scalable supervision introduces structural trade-offs that cannot be eliminated by pipeline improvements alone. These observations suggest that future progress will require advances beyond incremental improvements, including even stronger temporal reasoning, moving beyond supervision toward unsupervised approaches, uncertainty-aware association, and more tightly integrated multi-modal representations.

These open challenges motivate the future work discussed in the following section.

6.2. FUTURE WORK

The following section outlines potential future research directions, structured to reflect the major themes of this thesis. It begins with visual multi-object tracking, focusing on areas where improvements are expected in the near term. Multi-modality is then considered as a cross-cutting topic that applies throughout various aspects of perception. This is followed by a discussion on how data requirements for tracking may be addressed, as well as considerations for future datasets. The section concludes with a discussion of

broader developments in automated driving shaped by recent research trends.

VISUAL MULTI-OBJECT TRACKING

Multi-object tracking remains a rapidly advancing field. This section discusses potential future work, focusing on the tracking method families introduced in Chapter 2 while also highlighting opportunities that extend beyond them.

As one of the most widely used paradigms, **tracking-by-detection** will benefit from the steady improvement of object detectors and appearance embeddings. Future work will likely focus on richer temporal features, stronger integration of motion priors, and association strategies that replace IoU- or Kalman-based heuristics. Methods such as ByteTrack [209] and StrongSORT [41] have shown that even relatively simple refinements can yield significant benchmark gains. Building on this trend, hybrid systems that incorporate advanced memory banks or lightweight graph reasoning modules may emerge, enabling association to be preserved across extended temporal horizons. Joint detection-and-embedding models like JDE [186] remain attractive, though still lagging behind separated pipelines. Closing this gap will require stronger backbones and explicit occlusion reasoning.

Person re-identification methods form an important component within many tracking-by-detection approaches, where appearance embeddings support reliable association when motion cues are ambiguous or disrupted by occlusions. As stronger feature extractors and foundational models become available, more powerful appearance embeddings can be leveraged [26]. Transformer-based formulations that explicitly model parts and body regions hold potential, as highlighted by recent work [87, 133, 206, 219], which focus on auto-aligned parts, diverse embeddings, and attention-based part modeling.

While tracking-by-detection approaches have achieved strong results, they remain limited by reliance on heuristics for association, challenges with long occlusions, and lack of global temporal reasoning. **Transformer-based** methods directly address these issues by employing self- and cross-attention to learn spatio-temporal reasoning from data rather than relying on hand-crafted heuristics. This makes them more effective for temporal detection and re-identification through occlusions and fragmented trajectories. Trained on large datasets, they capture long-range dependencies and thus further enhance robustness. Three key challenges remain: First, high computational cost and limited inference speed, which hinder real-time deployment. Second, substantial data and resource requirements for training. And third, a persistent performance gap compared to strong tracking-by-detection baselines on public benchmarks. Future work will need to improve efficiency through deformable or hierarchical attention [96, 220], distillation [35, 53], or pruning strategies [80], and also make training more stable and effective. Recent advances such as query assignment refinements and query denoising [37, 45, 192] already point in this direction, making training more robust and improving performance. Ultimately, incremental progress is required to tackle the open points and close the gap to existing tracking baselines. Yet, because transformer-based methods do not easily accommodate heuristics and ad-hoc refinements, closing this gap on current benchmarks may prove difficult, again highlighting the importance of larger and more diverse datasets.

There is growing interest in **graph-based** formulations, with nodes representing detections and edges encoding potential associations. Classical graph approaches guaranteed globally optimal associations but relied heavily on heuristic cost functions (cf. Section 3.3). Recent advances with graph neural networks allow affinity metrics to be learned directly and enable joint reasoning across objects [138, 191]. This yields a powerful synergy: the GNN can weigh multiple cues (appearance, motion, spatial proximity) while an optimal solver (e.g., shortest paths or linear assignment) provides globally consistent tracks, as done in the recent work such as SSPGNN [49]. Future research will likely focus on making graph solvers fully differentiable so that the entire pipeline, from feature extraction to final associations, can be trained jointly on large datasets. Moreover, graph-based trackers open the door to explicitly modeling inter-object relations, e.g., by encoding pedestrians walking together or cyclists interacting with traffic as graph constraints.

Beyond these method families, several broader challenges remain unresolved. One major direction is the explicit use of **uncertainty**. Persistent occlusions and dense crowds frequently disrupt identity continuity, and recent work highlights the value of leveraging uncertainty rather than ignoring it. For instance, U²MOT [93] introduced an association-level uncertainty metric to detect uncertain matches and correct them, stabilizing unsupervised training and improving embeddings. This illustrates how uncertainty can be used to defer decisions, maintain multiple hypotheses, or re-label tracklets rather than prematurely committing to erroneous associations. Beyond unsupervised settings, future research may draw on Bayesian filtering ideas or probabilistic track representations, allowing trackers to carry distributions over identities or positions, thereby improving robustness without excessive complexity.

Finally, recent **generative approaches** provide new perspectives. Diffusion-based models [102, 104] treat associations as denoising problems, iteratively refining them into consistent tracks. Neural rendering methods [126] reframe tracking as inverse graphics problem by fitting generative 3D models across frames to preserve identities. Though computationally heavy, these techniques point toward trackers that reason through generative models of the world, offering richer cues for identity continuity and interpretability.

MULTI-MODAL DIRECTIONS IN FUTURE PERCEPTION

Multi-modal tracking extends visual pipelines by combining information from cameras with LiDAR [70, 177, 204, 207], radar [127, 204], or thermal sensors [221]. Multi-camera systems already demonstrate clear benefits [161], and future work will continue to shift toward 3D formulations and true multi-modal integrations. This is especially relevant in adverse conditions such as nighttime, which remains a major challenge as discussed in Chapter 1. Transformer-based methods are promising in this regard, since they model objects as tokens that naturally extend to multi-view and multi-modal reasoning. At the same time, established BEV-, voxel-, or full latent-based formulations provide strong baselines that can be further improved through tighter fusion and better exploitation of geometric cues.

Building on these developments in multi-modal tracking, identity modeling is also moving toward **multi-modal and cross-modal ReID** formulations. RGB-thermal [71,

[185, 206], RGB-depth [21, 52, 115], and LiDAR-based approaches [51, 168] offer robustness in diverse conditions. In addition, combining vision with language supervision opens up the possibility to exploit foundational models such as CLIP [137], allowing ReID systems to align appearance features with semantic descriptions and benefit from large-scale vision-language pre-training. To fully exploit such modalities, future systems require dedicated backbones trained jointly on multi-sensor data like shown in [85], rather than relying on image-pretrained networks. In parallel, ReID as a standalone component may become less prominent as tracking methods move toward transformer-based and end-to-end formulations, where identification is integrated more tightly into the overall tracking pipeline.

A related direction concerns the availability of **multi-modal datasets**, which offer synchronized sensor streams and richer scene context. As demonstrated by large-scale benchmarks such as Waymo [164] and nuScenes [14], datasets now routinely provide camera, LiDAR, and detailed map information. Going forward, radar and thermal sensors should also be incorporated, as seen in recent efforts such as the View of Delft dataset [128] for radar and the View of Delft Prediction dataset [9] for map. Map data, in particular, can be considered its own modality - while not a *sensor* modality - because it provides structured, high-level environmental information, that no onboard sensor directly observes. It captures road geometry, traffic infrastructure, lane topology, and permissible motion patterns, offering global context that complements local sensor measurements. Such multi-modal resources would allow evaluation under degraded visibility and in challenging conditions, and they also provide the essential data needed to begin developing and benchmarking new methods in the first place. This also implies the need for carefully designed annotation strategies across modalities, ranging from 2D image boxes to 3D cuboids, map-aligned object references, and potentially novel representations for radar or other sensor types.

SOLVING DATA-NEEDS FOR MULTI-OBJECT TRACKING

The semi-supervised trajectory generation pipeline introduced in Chapter 3 demonstrated that annotation costs can be significantly reduced, while providing useful supervision for training and evaluation. Future work may improve its components by integrating stronger detectors, potentially leveraging multiple sensors and temporal context, adopting more discriminative appearance models, and employing further developed offline tracking strategies to generate pseudo ground-truth. Yet, by design, this approach is limited in scalability due to its reliance on manual keyframes.

To overcome this scaling issue, **auto-labeling frameworks** can be used. Systems such as Auto4D [193] or off-board perception models such as [136] can generate large volumes of pseudo GT once trained on an initial dataset. In this sense, auto-labeling can be viewed as a semi-supervised approach, since it still depends on an initial supervised seed before large-scale label generation is possible. While such approaches scale more effectively to large amounts of raw data, they often operate with minimal direct human supervision, which can make them prone to systematic errors or overlooked edge cases. To safeguard quality, human-in-the-loop mechanisms are necessary. **Active learning** frameworks such as SPAMming Labels [19] offer a promising solution by focusing annotation effort only on ambiguous or uncertain cases, bringing annotation quality closer to full supervision while keeping costs low. Compared to the semi-supervised pipeline

presented in this thesis, these strategies appear more adaptive and scalable, showing strong potential for industrial-scale applications. At the same time, their dependence on substantial in-domain data for training robust baseline models, together with the integration of continuous human-in-the-loop mechanisms, remain open challenges that limit their immediate practicality.

Beyond semi-supervised approaches, **self-supervised and unsupervised tracking** approaches offer a path to bypass manual labels altogether. Unsupervised representation learning has already demonstrated strong results in vision, as seen in DINOv1 [18], and similar progress is emerging in 3D perception with unsupervised detection methods such as [83]. Unsupervised tracking methods can exploit inherent structure in videos, for example that a track forward and backward in time should lead to the same identity (temporal consistency), or that multiple augmented views of the same trajectory should yield consistent associations. Temporal consistency and path agreement losses are used by Lu et al. [98] to learn associations directly from unlabeled video. Uncertainty-aware unsupervised methods such as U²MOT [93] incorporate reliability metrics to detect and correct risky associations, stabilizing training without labels. Together, these approaches illustrate that large unlabeled data can be harnessed for tracking if the training objective is designed carefully.

Looking further ahead, one-shot and zero-shot tracking using foundation models could reduce the need for dataset-specific training altogether. By transferring general visual and semantic knowledge, such models could be prompted to track novel object classes or specific persons without task-specific supervision. This direction hints at a future where solving data scarcity relies not only on efficient labeling pipelines but also on leveraging broad pre-trained models.

NEXT GENERATION TRACKING DATASETS

The ECP2.0 dataset presented in Chapter 4 expanded both size and diversity and remains an important contribution of this thesis. Nonetheless, future datasets must broaden their scope considerably. A key improvement point is geographic and environmental diversity. While ECP2.0 already spans several European countries, questions of generalization across continents remain, and even larger-scale data might still be required as model complexity increases. Richer variety in training data is essential, capturing cultural differences in appearance and behavior, such as jaywalking norms, group sizes, or clothing styles, as well as environmental extremes including nighttime conditions, heavy rain, snow, or glare from sunrise and sunset. More data of this kind would not only strengthen training but also enable benchmarks to better highlight open issues within the tracking community and uncover potential generalization gaps.

Second, annotation quality and data curation remain crucial. As tracking methods improve, annotation errors - whether introduced by humans or semi-automatic tools - can become the limiting factor. This raises the need to revisit annotation practices, including systematic quality checks such as [132] or compensatory learning strategies that account for noise in ground-truth.

Lastly, it is not only about producing more data, but also about collecting the right data. Safety-critical long-tail events such as rare pedestrian behaviors, children suddenly entering the road, or complex cyclist maneuvers are still underrepresented. Addressing this requires large-scale data collection campaigns, combined with intelligent data

indexing and retrieval, as well as advances in simulation and generative modeling to synthesize realistic but uncommon scenarios - or scenarios which are too dangerous to record.

OUTLOOK: BEYOND CURRENT PARADIGMS

Foundation models are extremely large models pre-trained on vast datasets, providing highly transferable representations for a wide range of downstream tasks. Instead of training a tracker from scratch on a few hundred hours of driving data, these models leverage millions of internet images or videos to encode general visual knowledge. Examples include DINOv1-3 [18, 123, 155] and the Segment Anything Models SAM 1 [72] and 2 [142], with extensions such as Track Anything [195] demonstrating how segmentation and tracking can be combined effectively. For multi-object tracking, such models promise stronger detectors, more robust appearance embeddings, and even zero-shot or one-shot tracking, where novel **VRU** classes can be recognized without dataset specific training. They also highlight new opportunities for integrating universal features into detection or re-identification. Current limitations remain in terms of efficiency, resource demands, and limited multi-modal support.

The traditional modular pipeline - separating perception, prediction, and planning - outlined in Chapters 1 and Chapter 2 is increasingly challenged by fully **end-to-end systems** such as UniAD [67] and VAD [68]. These models unify the driving stack and allow gradients to propagate across the entire system, enabling joint optimization for safety and performance. In such settings, tracking becomes an implicit component that supports planning rather than being a standalone module. A tracking module inside such a stack might explicitly generate object trajectories, enabling evaluation, but it could also provide only latent information such as feature maps or tokens that downstream modules use. This shift reframes multi-object tracking: instead of producing isolated trajectories, it becomes a mechanism for providing reliable intermediate states that help downstream reasoning remain consistent and interpretable to enable safe driving.

A parallel line of work explores **world models** such as GAIA-1 [64], GAIA-2 [150], and DriveDreamer [183]. These generative approaches simulate entire driving environments, requiring them to maintain object identities over time, which embeds tracking naturally into their predictions. By producing coherent simulated futures, world models not only unify tracking and prediction but also enable large-scale synthetic data generation, particularly for rare or dangerous scenarios that are difficult to capture in reality. This integration blurs the line between perception, forecasting, and planning, situating multi-object tracking as one part of a broader generative understanding of traffic scenes.

Another promising frontier lies in **multi-modal foundation models** that integrate vision, language, and action. CLIP [137] exemplifies how linking image and text embeddings can provide open-vocabulary recognition. When combined with end-to-end pipelines, text prompts or language-grounded queries could condition perception and decision making, allowing systems to adapt to unseen situations. Early explorations such as DriveGPT-4 [190] highlight how vision-language models can describe driving scenes and suggest actions, pointing toward systems where tracking is coupled with semantic reasoning and interpretable decision support. The bigger picture here is a system that reasons about traffic in human-like terms: understanding not just where an object is,

but what it might do and why. By combining visual cues with action signals and language reasoning, such models could answer questions such as: *which pedestrian poses the greatest risk* or *what a pedestrian is likely intending to do next*? This level of reasoning leverages the vast behavioral knowledge encoded in large language models and has particular relevance for VRUs, whose behavior often defies rigid traffic rules and requires contextual, socially-aware interpretation.

Bringing all aspects together, the **long-term vision for automated driving perception** lies in integrated systems that unite end-to-end learning, foundation models, and continual self-improvement through simulation and real-world feedback. Such systems may eventually surpass human driving capabilities, limited mainly by available compute. Yet, this path remains challenging and will unfold incrementally. Within this broader journey, the contributions of this thesis represent a step toward more robust perception: demonstrating how large-scale trajectory data can be created semi-automatically, how multi-modal cues can enhance person re-identification, and how these insights strengthen the foundations for safe interaction with vulnerable road users.

ACKNOWLEDGEMENTS

While writing this dissertation I had the luck to meet many outstanding people, who actively or passively made this work possible. Here I want to mention a few special ones.

First and foremost, I want to express my *huge* gratitude to Prof. dr. Dariu M. Gavrila, my advisor and promoter. Thank you, Dariu, for a long journey marked by many valuable discussions, constructive feedback, and thoughtful supervision. Looking back, I am confident that this guidance did not only shape this dissertation, but also helped me grow into a better engineer and researcher.

Furthermore, I would like to express my appreciation to my second promoter Dr. Julian F.P. Kooij. Thank you for taking the time to review this thesis and for your constructive feedback, which significantly helped to improve its structure and clarity. I would also like to thank the members of the committee for their time and careful evaluation of this work.

Aside from my deep enthusiasm and interest in the topic itself, one of the main reasons for choosing the PhD position at Mercedes was the unique combination of an exciting research topic and an outstanding team with a high level of technical expertise. I would like to thank Dr. Ulrich Kreßel for his leadership, his constant support, and for always having an open ear when solutions were needed. I am also grateful to Prof. Dr. Fabian Flohr for introducing me to the field of computer vision and VRU tracking through my Master's thesis, which sparked my interest in the research direction of this thesis.

Many thanks go toward the other PhD students in our team: To Dr. Markus Braun: I could not have hoped for a better colleague to work alongside with than you. Thank you for the hundreds of hours of coffee talks, coding and paper sessions, and for making it possible to tackle an effort such as the ECP dataset together, including all the recording trips and accompanying fee issues. I always looked up to you - and still do - for your technical expertise and your approach to problem solving. As we are already at this first name, many thanks to Dr. Markus Roth. We not only shared an office, but also a close professional and personal connection over the past years. Your positive attitude consistently rubbed off on me and helped me to regain motivation and drive. Thanks to Julian Wiederer. We moved together to Stuttgart, and I always enjoyed our conversations and shared experiences along the way. Finishing our PhDs almost at the same time makes this journey feel even more special - and I still envy your remarkably structured way of managing time. Thanks also to Dr. Julian Mueller. We started our PhD journeys almost at the same time, although you were clearly the more efficient (and slightly "*luckier*") one. Thank you, Phillip Czech. The many hours we spent together in and outside of the office were some of the most enjoyable parts of everyday work, and your energy always lifted the mood. Thanks for pushing the ECP data forward with so much drive, and for motivating me time and time again to head to the gym. Lastly, I would like to thank Marius Daehling. While officially supervised by me, working together has been equally

enriching. You are well on your way to your own PhD, and I am very happy to support and supervise you on that path.

Overall, I am happy to say that all those mentioned here, Markus, Markus, Julian, Julian, Phillip, and Marius, are not just "colleagues" who made life during writing this thesis easier, but true friends which I am deeply happy for. Beyond this close circle, I would also like to thank my colleagues at Mercedes-Benz more broadly, especially Marius Cordts, my current team lead, Bharanidhar Duraisamy, and Michael Gabb and Otto Loehlein, who supervised me when I first started as a working student. I would also like to thank my colleagues at TU Delft, especially Ronald Ensing, for the inspiring discussions and technical collaboration along the way.

I was fortunate to have many friends who made the time during my PhD not only easier, but also more enjoyable, and who gave me energy to keep going. Special thanks to Kim, for many great conversations and especially for your help with the design of this front cover. The thesis would look a lot more boring without you. I would also like to thank Mark, one of my closest friends. Sharing the PhD journey with you, still ongoing and perhaps a bit overdue for both of us, made many things easier and more relatable. In theory, you also brought me to Mercedes in the first place. And whenever motivation was running low, reminding each other that "it do be like that sometimes - but not all the times" helped more than we might admit. Keep rolling Marshmallow.

This thesis, and my whole career, would not have been possible without the unwavering support of my family. I would like to especially thank my parents Silke and Gerhard, as well as my sister, Stephanie. I am deeply grateful for the freedom, trust, and encouragement my parents gave me throughout my life. Their support, particularly in allowing and encouraging me to study computer science, made this dissertation possible in the first place. You always had my back, encouraged me during difficult times, listened when things were challenging, and gave me the strength to continue. I could not have wished for more loving and supportive parents. Thank you from the bottom of my heart.

Lastly, and with my deepest gratitude, I want to thank my girlfriend Paula. Without your constant support, motivation, and belief in me, this thesis would not have been finished. You always picked me up when things were difficult and motivation was lacking, challenged me when needed, and celebrated every success - often even more enthusiastically than I did myself. Thank you for always giving me the space and support to work on this thesis, for the technical discussions and solution-space talks, and for all the late-night sandwich days. I am truly looking forward to the time after this thesis and to everything that lies ahead for us together.

CURRICULUM VITÆ

Sebastian Alexander KREBS

26-04-1992 Born in Ulm, Germany.

EDUCATION

2016-Present Ph.D. in the Cognitive Robotics department
Delft University of Technology, Netherlands
Thesis: Visual Tracking of Vulnerable Road Users for Automated
Driving

2013–2016 M.Sc. Computer Science
University of Ulm, Germany

2010–2013 B.Sc. Computer Science
Ulm University of Applied Sciences, Germany

PROFESSIONAL EXPERIENCE

2016–Present Mercedes-Benz AG, Germany
Machine Learning Engineer in the department *Perception and Maps*.

LIST OF PUBLICATIONS

THESIS PUBLICATIONS

6. **S. Krebs**, D. M. Gavrila “Camera- and LiDAR-based Person Re-identification”, [Proc. of the IEEE Intelligent Vehicles Symposium \(IV\)](#), 2025, pp. 1408–1414.

Author contributions: S. Krebs implemented and evaluated the proposed method. S. Krebs wrote the paper. D. M. Gavrila provided guidance and supervision.

5. **S. Krebs**, M. Braun, and D. M. Gavrila “EuroCity Persons 2.0: A Large and Diverse Dataset of Persons in Traffic”, [IEEE Trans. on Pattern Analysis and Machine Intelligence \(TPAMI\)](#), 2024, vol. 46, no. 12, pp. 10929–10943.

Author contributions: S. Krebs implemented and evaluated the proposed method for the creation of the ECP2.0 dataset. S. Krebs wrote the article with the help of D. M. Gavrila. S. Krebs and M. Braun built up the sensor vehicle and recorded the data during several journeys. D. M. Gavrila provided guidance and supervision.

4. M. Braun, **S. Krebs**, and D. M. Gavrila “ECP2.5D - Person Localization in Traffic Scenes”, [Proc. of the IEEE Intelligent Vehicles Symposium \(IV\)](#), 2020, pp. 1694–1701.

Author contributions: Equal contribution from the first two authors. S. Krebs and M. Braun created and implemented the uplifting method used to create the ECP2.5D dataset, S. Krebs implemented the ego-motion correction of LiDAR point clouds and the clustering algorithm. M. Braun implemented and evaluated the proposed baseline method for monocular 2.5D localization. M. Braun and S. Krebs wrote the paper, D. M. Gavrila provided guidance and supervision.

3. **S. Krebs**, M. Braun, and D. M. Gavrila “Generating 3D Person Trajectories from Sparse Image Annotations in an Intelligent Vehicles Setting”, [Proc. of the IEEE Intelligent Transportation Systems Conf. \(ITSC\)](#), 2019, pp. 783–788.

Author contributions: S. Krebs designed, implemented, and evaluated the proposed system, and wrote the paper, M. Braun provided technical support and feedback, D. M. Gavrila provided guidance and supervision.

2. M. Braun, **S. Krebs**, F. Flohr and D. M. Gavrila “EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes”, [IEEE Trans. on Pattern Analysis and Machine Intelligence \(TPAMI\)](#), 2019, vol. 41, no. 8, pp. 1844–1861.

Author contributions: M. Braun implemented, modified, and evaluated the benchmark methods and experiments. S. Krebs, M. Braun, and F. Flohr built up the sensor vehicle and recorded the data during several journeys. S. Krebs helped M. Braun to create the detection benchmark with F. Flohr. F. Flohr implemented the server framework to be used for online benchmarking. S. Krebs helped M. Braun in writing the journal article along with D. M. Gavrila. D. M. Gavrila furthermore provided guidance and supervision.

1. **S. Krebs**, B. Duraisamy, and F. Flohr “A survey on leveraging deep neural networks for object tracking”, *Proc. of the IEEE Intelligent Transportation Systems Conf. Workshops (ITSC)*, 2017, pp. 411–418.

Author contributions: S. Krebs conducted the literature review, structured the survey, analyzed the state-of-the-art, and wrote the paper. B. Duraisamy and F. Flohr provided guidance and supervision.

OTHER PUBLICATIONS

3. M. Braun, F. Flohr, **S. Krebs**, U. Kreßel, D. M. Gavrila “Simple Pair Pose - Pairwise Human Pose Estimation in Dense Urban Traffic Scenes”, *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 1545–1552.
2. I. Cieslik, J. Kovaceva, M. Bruyas, D. Large, M. Kunert, **S. Krebs**, M. Arbritmann “Improving the effectiveness of active safety systems to significantly reduce accidents with vulnerable road users - the Project PROSPECT (Proactive Safety for Pedestrians and Cyclists)”, *International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2019 .
1. A. Fregin, M. Roth, M. Braun, **S. Krebs**, and F. Flohr, “Building a Computer Vision Research Vehicle with ROS”, *Proc. of the ROSCon*, 2017.

PATENTS

3. M. Braun, **S. Krebs**, “Langstrecken-Detektion ohne Langstrecken-Annotationen”, Submitted as German Patent, 2025.
2. N. Schneider, **S. Krebs**, B. Chen, “Camera 3D Obstacle Detection Suppression Based on 3D Cuboid Features”, Submitted as Patent in the United Kingdom, 2025.
1. J. Wiederer, **S. Krebs**, P. Czech, “System zur Automatischen 3D Posenbestimmung im Kamerabild eines Bewegten Fahrzeugs durch Kombination eines GNSS und Motion Capture Systems”, German Patent No. DE102022001327, 2022.

BIBLIOGRAPHY

- [1] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky. “BoT-SORT: Robust Associations Multi-Pedestrian Tracking”. In: *arXiv preprint arXiv:2206.14651* (2022).
- [2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. “Social LSTM: Human Trajectory Prediction in Crowded Spaces”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 961–971.
- [3] A. Andriyenko, K. Schindler, and S. Roth. “Discrete-continuous optimization for multi-target tracking”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 1926–1933.
- [4] Y. Bar-Shalom, F. Daum, and J. Huang. “The Probabilistic Data Association Filter”. In: *IEEE Control Systems Magazine*: December (2009), pp. 82–100.
- [5] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. “Multiple object tracking using k-shortest paths optimization”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 33:9 (2011), pp. 1806–1819.
- [6] P. Bergmann, T. Meinhardt, and L. Leal-Taixe. “Tracking Without Bells and Whistles”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2019, pp. 941–951.
- [7] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. “Simple Online and Realtime Tracking”. In: *Proc. of the International Conf. on Image Processing (ICIP)*. 2016, pp. 3464–3468.
- [8] S. Blackman. “Multiple hypothesis tracking for multiple target tracking”. English. In: *IEEE Aerospace and Electronic Systems Magazine* 19:1 (2004), pp. 5–18.
- [9] H. J. Boekema, B. K. Martens, J. F. P. Kooij, and D. M. Gavrila. “Multi-Class Trajectory Prediction in Urban Traffic Using the View-of-Delft Prediction Dataset”. In: *IEEE Robotics and Automation Letters (RA-L)* 9:5 (2024), pp. 4806–4813.
- [10] A. Boragule, H. Jang, N. Ha, and M. Jeon. “Pixel-Guided Association for Multi-Object Tracking”. In: *Sensors* 22:22 (2022), p. 8922.
- [11] M. Braun, S. Krebs, and D. M. Gavrila. “ECP2.5D - Person Localization in Traffic Scenes”. In: *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*. 2020, pp. 1694–1701.
- [12] M. Braun, F. B. Flohr, S. Krebs, U. Kreßel, and D. M. Gavrila. “Simple Pair Pose - Pairwise Human Pose Estimation in Dense Urban Traffic Scenes”. In: *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*. 2021, pp. 1545–1552.
- [13] M. Braun, S. Krebs, F. B. Flohr, and D. M. Gavrila. “EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 41:8 (2019), pp. 1844–1861.

- [14] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. “nuScenes: A Multimodal Dataset for Autonomous Driving”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11621–11631.
- [15] J. Cai, M. Xu, W. Li, Y. Xiong, W. Xia, Z. Tu, and S. Soatto. “MeMOT: Multi-Object Tracking With Memory”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 8090–8100.
- [16] J. Cao, J. Pang, X. Weng, R. Khrodgar, and K. Kitani. “Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 9686–9696.
- [17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. “End-to-End Object Detection with Transformers”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 213–229.
- [18] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. “Emerging Properties in Self-Supervised Vision Transformers”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2021, pp. 9630–9640.
- [19] O. Cetintas, T. Meinhardt, G. Brasó, and L. Leal-Taixé. “SPAMming Labels: Efficient Annotations for the Trackers of Tomorrow”. In: *European Conference on Computer Vision (ECCV)*. Vol. 15070. 2025, pp. 377–395.
- [20] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays. “Argoverse: 3D Tracking and Forecasting With Rich Maps”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 8748–8757.
- [21] A. Chavan et al. “Towards Global Localization Using Multi-Modal Object-Instance Re-Identification”. In: *arXiv preprint arXiv:2409.12002* (2024).
- [22] C. Chen, L. Zanotti Fragonara, and A. Tsourdos. “Relation3DMOT: Exploiting Deep Affinity for 3D Multi-Object Tracking from View Aggregation”. In: *Sensors* 21:6 (2021), p. 2113.
- [23] C.-F. R. Chen, Q. Fan, and R. Panda. “Crossvit: Cross-attention multi-scale vision transformer for image classification”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2021, pp. 357–366.
- [24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”. In: *European Conference on Computer Vision (ECCV)*. 2018.
- [25] P. Chen et al. “Occlude Them All: Occlusion-Aware Attention Network for Occluded Person Re-ID”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2021, pp. 11833–11842.
- [26] 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: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 15050–15061.

- [27] Z. Chi, H. Li, H. Lu, and M.-H. Yang. “Dual Deep Network for Visual Tracking”. In: *IEEE Transactions on Image Processing* (2017), pp. 2005–2015.
- [28] S. Chopra, R. Hadsell, and L. Y. “Learning a similarity metric discriminatively, with application to face verification”. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2005), pp. 349–356.
- [29] J. B. Cicchino. “Effects of Automatic Emergency Braking Systems on Pedestrian Crash Risk”. In: *Accident Analysis & Prevention* 172 (2022), p. 106686.
- [30] I. Cieslik, J. Kovaceva, M.-P. Bruyas, D. Large, M. Kunert, S. Krebs, and M. Arbitmann. “Improving the effectiveness of active safety systems to significantly reduce accidents with vulnerable road users - the Project PROSPECT (Proactive Safety for Pedestrians and Cyclists)”. In: *International Technical Conference on the Enhanced Safety of Vehicles (ESV)*. 2019.
- [31] D. Clark, K. Panta, and B.-n. Vo. “The GM-PHD Filter Multiple Target Tracker”. English. In: *International Conference on Information Fusion (FUSION)*. 2006, pp. 1–8.
- [32] P. Cong, X. Zhu, F. Qiao, Y. Ren, X. Peng, Y. Hou, L. Xu, R. Yang, D. Manocha, and Y. Ma. “STCrowd: A Multimodal Dataset for Pedestrian Perception in Crowded Scenes”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 19608–19617.
- [33] M. Cordts et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016.
- [34] O. de Groot, A. Bertipaglia, H. Boekema, V. Jain, M. Kegl, V. Kotian, T. Lentsch, Y. Lin, C. Messiou, E. Schippers, F. Tajdari, S. Wang, Z. Xia, M. Zaffar, R. Ensing, M. Garzon, J. Alonso-Mora, H. Caesar, L. Ferranti, R. Happee, J. F. P. Kooij, G. Papaioannou, B. Shyrokau, and D. M. Gavrila. “A Vehicle System for Navigating Among Vulnerable Road Users Including Remote Operation”. In: *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*. 2025, pp. 2482–2489.
- [35] P. De Rijk, L. Schneider, M. Cordts, and D. M. Gavrila. “Structural Knowledge Distillation for Object Detection”. In: *Adv. in Neural Information Processing Systems (NeurIPS)*. Vol. 35. 2022, pp. 3858–3870.
- [36] J. Dequaire, D. Rao, P. Ondruska, D. Wang, and I. Posner. “Deep Tracking on the Move: Learning to Track the World from a Moving Vehicle using Recurrent Neural Networks”. In: *arXiv preprint arXiv:1609.09365* (2016).
- [37] S. Ding, Y. Yang, J. Wiederer, M. Braun, P. Li, J. Gall, and B. Yang. “TQD-Track: Temporal Query Denoising for 3D Multi-Object Tracking”. In: *arXiv preprint arXiv:2504.03258* (2025).
- [38] P. Dollár, C. Wojek, B. Schiele, and P. Perona. “Pedestrian Detection: A Benchmark”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2009.

- [39] P. Dollár, C. Wojek, B. Schiele, and P. Perona. “Pedestrian Detection: An Evaluation of the State of the Art”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 34 (2012).
- [40] A. Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
- [41] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng. “StrongSORT: Make DeepSORT Great Again”. In: *IEEE Transactions on Multimedia* 25 (2023), pp. 8725–8737.
- [42] European Commission. *Mobility & Transport - Road Safety: Facts and Figures 2024 (2022 data)*. Tech. rep. European Road Safety Observatory, 2024.
- [43] European Commission. *Road safety thematic report - Pedestrians*. Tech. rep. European Road Safety Observatory, 2023.
- [44] European Commission and European Climate, Infrastructure and Environment Executive Agency. *EU road safety - Towards “Vision Zero”*. Tech. rep. Publications Office of the European Union, 2022.
- [45] T. Fu, X. Wang, H. Yu, K. Niu, B. Li, and X. Xue. “DeNoising-MOT: Towards Multiple Object Tracking with Severe Occlusions”. In: *Proc. of the ACM International Conference on Multimedia*. 2023, pp. 2734–2743.
- [46] Q. Gan, Q. Guo, Z. Zhang, and K. Cho. “First Step toward Model-Free, Anonymous Object Tracking with Recurrent Neural Networks”. In: *arXiv preprint arXiv:1511.06425* (2015).
- [47] A. Geiger, P. Lenz, and R. Urtasun. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 3354–3361.
- [48] P. Gil-Jiménez, H. Gómez-Moreno, R. López-Sastre, and S. Maldonado-Bascón. “Geometric bounding box interpolation: an alternative for efficient video annotation”. In: *EURASIP Journal on Image and Video Processing* 2016:1 (2016), p. 8.
- [49] G. Golias, M. Nakura-Fan, and V. Ablavsky. “SSP-GNN: Learning to Track via Bilevel Optimization”. In: *International Conference on Information Fusion (FUSION)*. 2024, pp. 1–8.
- [50] L. Gu, S. Sun, X. Liu, and X. Li. “CenterTrack3D: Improved CenterTrack More Suitable for Three-Dimensional Objects”. In: *Journal of Autonomous Vehicles and Systems* 1:2 (2021), p. 021004.
- [51] W. Guo et al. “LiDAR-based Person Re-identification”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 17437–17447.
- [52] Hafner, Frank M., Bhuyian, Amran, Kooij, Julian F. P., and Granger, Eric. “Cross-Modal Distillation for RGB-depth Person Re-Identification”. In: *Computer Vision and Image Understanding* 216 (2022), p. 103352.
- [53] Z. Hao, J. Guo, D. Jia, K. Han, Y. Tang, C. Zhang, H. Hu, and Y. Wang. “Learning Efficient Vision Transformers via Fine-Grained Manifold Distillation”. In: *Adv. in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 9164–9175.

- [54] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao. “Generalizable Pedestrian Detection: The Elephant in the Room”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11328–11337.
- [55] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.
- [56] L. He et al. “Fastreid: A pytorch toolbox for general instance re-identification”. In: *Proc. of the ACM International Conference on Multimedia*. 2023, pp. 9664–9667.
- [57] S. He et al. “TransReID: Transformer-Based Object Re-Identification”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2021, pp. 15013–15022.
- [58] D. Held, S. Thrun, and S. Savarese. “Learning to Track at 100 FPS with Deep Regression Networks”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 749–765.
- [59] A. Hermans, L. Beyer, and B. Leibe. “In Defense of the Triplet Loss for Person Re-Identification”. In: *arXiv preprint arXiv:1703.07737* (2017).
- [60] S. Hoermann, M. Bach, and K. Dietmayer. “Dynamic Occupancy Grid Prediction for Urban Autonomous Driving: A Deep Learning Approach with Fully Automatic Labeling”. In: *Proc. of the International Conf. on Robotics and Automation (ICRA)*. 2018, pp. 2056–2063.
- [61] S. Hong, T. You, S. Kwak, and B. Han. “Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network”. In: *Proc. of the International Conf. on Machine Learning (ICML)*. 2015, pp. 597–606.
- [62] H. Hou, C. Shen, X. Zhang, and W. Gao. “CSMOT: Make One-Shot Multi-Object Tracking in Crowded Scenes Great Again”. In: *Sensors* 23:7 (2023), p. 3782.
- [63] R. Hou et al. “VRSTC: Occlusion-Free Video Person Re-Identification”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 7183–7192.
- [64] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado. “GAIA-1: A Generative World Model for Autonomous Driving”. In: *arXiv preprint arXiv:2309.17080* (2023).
- [65] H.-N. Hu, Y.-H. Yang, T. Fischer, T. Darrell, F. Yu, and M. Sun. “Monocular Quasi-Dense 3D Object Tracking”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 45:2 (2023), pp. 1992–2008.
- [66] Y. Hu, A. Niu, Y. Zhu, Q. Yan, J. Sun, and Y. Zhang. “Multiple Object Tracking Based on Occlusion-Aware Embedding Consistency Learning”. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 9521–9525.
- [67] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li. “Planning-Oriented Autonomous Driving”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 17853–17862.

- [68] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang. "VAD: Vectorized Scene Representation for Efficient Autonomous Driving". In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2023, pp. 8306–8316.
- [69] R. E. Kalman. "A new approach to linear filtering and prediction problems". In: *Journal of basic Engineering* 82 (1960), pp. 35–45.
- [70] A. Kim, A. Ošep, and L. Leal-Taixé. "EagerMOT: 3D Multi-Object Tracking via Sensor Fusion". In: *Proc. of the International Conf. on Robotics and Automation (ICRA)*. 2021, pp. 11315–11321.
- [71] S. Kim, S. Gwon, and K. Seo. "Enhancing Diverse Intra-identity Representation for Visible-Infrared Person Re-Identification". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2024, pp. 2501–2510.
- [72] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. "Segment Anything". In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2023, pp. 3992–4003.
- [73] S. Krebs, M. Braun, and D. M. Gavrila. "EuroCity Persons 2.0: A Large and Diverse Dataset of Persons in Traffic". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 46:12 (2024), pp. 10929–10943.
- [74] S. Krebs, M. Braun, and D. M. Gavrila. "Generating 3D Person Trajectories from Sparse Image Annotations in an Intelligent Vehicles Setting". In: *Proc. of the IEEE Intelligent Transportation Systems Conf. (ITSC)*. 2019, pp. 783–788.
- [75] S. Krebs, B. Duraisamy, and F. Flohr. "A survey on leveraging deep neural networks for object tracking". In: *Proc. of the IEEE Intelligent Transportation Systems Conf. Workshops (ITSC)*. 2017, pp. 411–418.
- [76] S. Krebs and D. M. Gavrila. "Camera- and LiDAR-based Person Re-identification". In: *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*. 2025, pp. 1408–1414.
- [77] H. W. Kuhn. "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2:1-2 (1955), pp. 83–97.
- [78] A. Kullgren, K. Amin, and C. Tingvall. "Effects on Crash Risk of Automatic Emergency Braking Systems for Pedestrians and Bicyclists". In: *Traffic Injury Prevention* 24:sup1 (2023), S111–S115.
- [79] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, et al. "The open images dataset v4". In: *International Journal of Computer Vision* 128:7 (2020), pp. 1956–1981.
- [80] W. Kwon, S. Kim, M. W. Mahoney, J. Hassoun, K. Keutzer, and A. Gholami. "A Fast Post-Training Pruning Framework for Transformers". In: *Adv. in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 24101–24116.
- [81] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler. "Learning by tracking: Siamese CNN for robust target association". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPR)*. 2016, pp. 33–40.

- [82] C. Lee and D. Hong. “Feature extraction using the Bhattacharyya distance”. In: *IEEE Transactions on Systems, Man, and Cybernetics*. Vol. 3. 1997, 2147–2150 vol.3.
- [83] T. Lentsch, H. Caesar, and D. M. Gavrilu. “Union: Unsupervised 3d Object Detection Using Object Appearance-Based Pseudo-Classes”. In: *Adv. in Neural Information Processing Systems (NeurIPS)* 37 (2024), pp. 22028–22046.
- [84] B. Li, C. Yang, and G. Xu. “Multi-pedestrian tracking based on feature learning method with lateral inhibition”. In: *IEEE International Conference on Information and Automation (ICIA)*. 2015.
- [85] H. Li, M. Ye, M. Zhang, and B. Du. “All in One Framework for Multimodal Re-Identification in the Wild”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 17459–17469.
- [86] K. Li, H. Rezatofighi, and I. Reid. “MOLTR: Multiple Object Localization, Tracking and Reconstruction From Monocular RGB Videos”. In: *IEEE Robotics and Automation Letters (RA-L)* 6:2 (2021), pp. 3341–3348.
- [87] W. Li et al. “DC-Former: Diverse and compact transformer for person re-identification”. In: *Proc. of the AAAI Conference on Artificial Intelligence*. Vol. 37. 2. 2023, pp. 1415–1423.
- [88] X. R. Li and V. P. Jilkov. “Survey of maneuvering targettracking . part I: dynamic models”. In: *IEEE Transactions on Aerospace and Electronic Systems* 39:4 (2003), pp. 1333–1364.
- [89] Z. Li, Y. Cui, Y. Lin, and Z. Fang. “MMF-Track: Multi-modal Multi-level Fusion for 3D Single Object Tracking”. In: *IEEE Transactions on Intelligent Vehicles (T-IV)* 9:1 (2024), pp. 1817–1829.
- [90] D. Liciotti et al. “Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration”. In: *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*. 2017, pp. 1–11.
- [91] B. Liu, E. Adeli, Z. Cao, K. H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles. “Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction”. In: *IEEE Robotics and Automation Letters (RA-L)* 5:2 (2020), pp. 3485–3492.
- [92] H. Liu, L. Hu, and L. Ma. “Online RGB-D Person Re-Identification Based on Metric Model Update”. In: *CAAI Transactions on Intelligence Technology* 2:1 (2017), pp. 48–55.
- [93] K. Liu, S. Jin, Z. Fu, Z. Chen, R. Jiang, and J. Ye. “Uncertainty-Aware Unsupervised Multi-Object Tracking”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2023, pp. 9996–10005.
- [94] Q. Liu, D. Chen, Q. Chu, L. Yuan, B. Liu, L. Zhang, and N. Yu. “Online Multi-Object Tracking with Unsupervised Re-Identification Learning and Occlusion Estimation”. In: *Neurocomputing* 483 (2022), pp. 333–347.
- [95] W. Liu, Y. Li, J. Wang, Z. Miao, and H. Qiu. “Multi-Object Tracking with Noisy Labels”. In: *IEEE Int. Conf. on Pattern Recognition and Artificial Intelligence (PRAI)*. 2022, pp. 443–449.

- [96] 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: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2021, pp. 10012–10022.
- [97] Z. Liu, X. Wang, C. Wang, W. Liu, and X. Bai. "SparseTrack: Multi-Object Tracking by Performing Scene Decomposition Based on Pseudo-Depth". In: *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* 35:5 (2025), pp. 4870–4882.
- [98] Z. Lu, B. Shuai, Y. Chen, Z. Xu, and D. Modolo. "Self-Supervised Multi-Object Tracking with Path Consistency". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 19016–19026.
- [99] 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". In: *International Journal of Computer Vision* 129:2 (2020), pp. 1–31.
- [100] J. Luiten, I. E. Zulfikar, and B. Leibe. "UnOVOST: Unsupervised Offline Video Object Segmentation and Tracking". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 2000–2009.
- [101] H. Luo et al. "Bag of tricks and a strong baseline for deep person re-identification". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPR)*. 2019.
- [102] R. Luo, Z. Song, L. Ma, J. Wei, W. Yang, and M. Yang. "DiffusionTrack: Diffusion Model for Multi-Object Tracking". In: *Proc. of the AAAI Conference on Artificial Intelligence* 38:5 (2024), pp. 3991–3999.
- [103] W. Luo et al. "Multiple Object Tracking: A Literature Review". In: *The journal of Artificial Intelligence (AIJ)* 293 (2021), p. 103448.
- [104] 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: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 19321–19330.
- [105] C. Ma, J. B. Huang, X. Yang, and M. H. Yang. "Hierarchical convolutional features for visual tracking". In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2016, pp. 3074–3082.
- [106] G. Maggolino, A. Ahmad, J. Cao, and K. Kitani. "Deep OC-Sort: Multi-Pedestrian Tracking by Adaptive Re-Identification". In: *Proc. of the International Conf. on Image Processing (ICIP)*. 2023, pp. 3025–3029.
- [107] S. Malla, B. Dariush, and C. Choi. "Titan: Future forecast using action priors". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 11183–11193.
- [108] S. Manen, M. Gygli, D. Dai, and L. Van Gool. "PathTrack: Fast Trajectory Annotation with Path Supervision". In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2017, pp. 290–299.

- [109] N. Marinello, M. Proesmans, and L. Van Gool. “TripletTrack: 3D Object Tracking Using Triplet Embeddings and LSTM”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 4500–4510.
- [110] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer. “TrackFormer: Multi-Object Tracking with Transformers”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 8834–8844.
- [111] D. Mihalcik and D. Doermann. “The design and implementation of ViPER”. In: *University of Maryland* (2003), pp. 234–241.
- [112] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. “MOT16: A Benchmark for Multi-Object Tracking”. In: *arXiv preprint arXiv:1603.00831* (2016).
- [113] A. Milan, S. Roth, and K. Schindler. “Continuous energy minimization for multitarget tracking.” In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 36:1 (2014), pp. 58–72.
- [114] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler. “Online Multi-Target Tracking Using Recurrent Neural Networks”. In: vol. 31. 1. 2017.
- [115] H. Mukhtar and M. U. G. Khan. “CMOT: A Cross-Modality Transformer for RGB-D Fusion in Person Re-Identification with Online Learning Capabilities”. In: *Knowledge-Based Systems* (2024).
- [116] M. Munaro et al. “One-Shot Person Re-identification with a Consumer Depth Camera”. In: *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition (ACVPR). Springer, 2014, pp. 161–181.
- [117] H. Nam and B. Han. “Learning Multi-domain Convolutional Neural Networks for Visual Tracking”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4293–4302.
- [118] G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, and Z. He. “Spatially supervised recurrent convolutional neural networks for visual object tracking”. In: *IEEE International Symposium on Circuits and Systems (ISCAS)*. 2017, pp. 1–4.
- [119] J. Niño-Castañeda, A. Frías-Velázquez, N. Bo Bo, M. Slembrouck, J. Guan, G. Debarb, B. Vanrumste, T. Tuytelaars, and W. Philips. “Scalable Semi-Automatic Annotation for Multi-Camera Person Tracking”. In: *IEEE Transactions on Image Processing* 25:5 (2016), pp. 2259–2274.
- [120] P. Ondruska, J. Dequaire, D. Z. Wang, and I. Posner. “End-to-End Tracking and Semantic Segmentation Using Recurrent Neural Networks”. In: *arXiv preprint arXiv:1604.05091* (2016).
- [121] P. Ondruska and I. Posner. “Deep tracking: Seeing beyond seeing using recurrent neural networks”. In: *Proc. of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016.
- [122] M. Onis. “WHO Child Growth Standards based on length/height, weight and age: WHO Child Growth Standards”. In: *Acta Paediatrica - ACTA PAEDIAT* (2007).

- [123] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. "DINOv2: Learning Robust Visual Features without Supervision". In: *Transactions on Machine Learning Research* (2024).
- [124] W. H. Organization. *Global status report on road safety 2018*. World Health Organization, 2018.
- [125] W. H. Organization. *Global status report on road safety 2023*. World Health Organization, 2023.
- [126] J. Ost, T. Banerjee, M. Bijelic, and F. Heide. "Inverse Neural Rendering for Explainable Multi-Object Tracking". In: *arXiv preprint arXiv:2404.12359* (2024).
- [127] A. Palfy, J. F. P. Kooij, and D. M. Gavrila. "Detecting Darting Out Pedestrians With Occlusion Aware Sensor Fusion of Radar and Stereo Camera". In: *IEEE Transactions on Intelligent Vehicles (T-IV)* 8:2 (2023), pp. 1459–1472.
- [128] A. Palfy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila. "Multi-Class Road User Detection With 3+1D Radar in the View-of-Delft Dataset". In: *IEEE Robotics and Automation Letters (RA-L)* 7:2 (2022), pp. 4961–4968.
- [129] X. Pan et al. "Two at once: Enhancing learning and generalization capacities via ibn-net". In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 464–479.
- [130] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu. "Quasi-Dense Similarity Learning for Multiple Object Tracking". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 164–173.
- [131] C. Patruno et al. "People Re-Identification Using Skeleton Standard Posture and Color Descriptors from RGB-D Data". In: *Pattern Recognition* 89 (2019), pp. 77–90.
- [132] S. Penquitt, J. Klees, R. Cakaj, D. Kondermann, M. Rottmann, and L. Schmarje. "From Label Error Detection to Correction: A Modular Framework and Benchmark for Object Detection Datasets". In: *arXiv preprint arXiv:2508.06556* (2025).
- [133] N. Pervaiz, M. M. Fraz, and M. Shahzad. "Per-Former: Rethinking Person Re-Identification Using Transformer Augmented with Self-Attention and Contextual Mapping". In: *The Visual Computer* (2023), pp. 4087–4102.
- [134] N. Phan et al. "LoGoViT: Local-global vision transformer for object re-identification". In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5.
- [135] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. "Globally-optimal greedy algorithms for tracking a variable number of objects". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2011, pp. 1201–1208.
- [136] C. R. Qi, Y. Zhou, M. Najibi, P. Sun, K. Vo, B. Deng, and D. Anguelov. "Offboard 3d object detection from point cloud sequences". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 6134–6144.

- [137] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proc. of the International Conf. on Machine Learning (ICML)*. 2021, pp. 8748–8763.
- [138] A. Rangesh, P. Maheshwari, M. Gebre, S. Mhatre, V. Ramezani, and M. M. Trivedi. “TrackMPNN: A Message Passing Graph Neural Architecture for Multi-Object Tracking”. In: *arXiv preprint arXiv:2101.04206* (2021).
- [139] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos. “Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2019, pp. 6262–6271.
- [140] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. “Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior”. In: *Proc. of the International Conf. on Computer Vision Workshops (ICCV Workshops)*. 2017, pp. 206–213.
- [141] H. E. Rauch, F. Tung, and C. T. Striebel. “Maximum likelihood estimates of linear dynamic systems”. In: *AIAA Journal* 3:8 (1965), pp. 1445–1450.
- [142] N. Ravi, V. Gabeur, Y.-T. 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.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer. “SAM 2: Segment Anything in Images and Videos”. In: *International Conference on Learning Representations (ICLR)*. 2025.
- [143] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. “You only look once: Unified, real-time object detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788.
- [144] S. Ren, K. He, R. Girshick, and J. Sun. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *Adv. in Neural Information Processing Systems (NeurIPS)*. 2015, pp. 91–99.
- [145] S. Reuter, Ba-Tuong Vo, Ba-Ngu Vo, and K. Dietmayer. “The Labeled Multi-Bernoulli Filter”. In: *IEEE Transactions on Signal Processing* 62:12 (2014), pp. 3246–3260.
- [146] On-Road Automated Driving (ORAD) Committee. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Tech. rep. SAE International, 2021, pp. 1–41.
- [147] M. Roth, J. Stapel, R. Happee, and D. M. Gavrila. “Driver and Pedestrian Mutual Awareness for Path Prediction and Collision Risk Estimation”. In: *IEEE Transactions on Intelligent Vehicles (T-IV)* 7:4 (2022), pp. 896–907.
- [148] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. “Human Motion Trajectory Prediction: A Survey”. In: *International Journal of Robotics Research (IJRR)* 39:8 (2020), pp. 895–935.
- [149] H. Ruff and M. Rothbart. *Attention in Early Development: Themes and Variations*. Oxford University Press, 2001, pp. 1–310.

- [150] L. Russell, A. Hu, L. Bertoni, G. Fedoseev, J. Shotton, E. Arani, and G. Corrado. “GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving”. In: *arXiv preprint arXiv:2503.20523* (2025).
- [151] V. S. Saravananarajan, R.-C. Chen, and L.-S. Chen. “LiDAR Point Cloud Data Processing in Autonomous Vehicles”. In: *International Conference on Electronics, Communication and Computing Technologies (ICECCT)*. 2021, pp. 1–5.
- [152] J. Seidenschwarz et al. “Simple Cues Lead to a Strong Multi-Object Tracker”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 13813–13823.
- [153] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua. “Tracking Multiple People under Global Appearance Constraints”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2011, pp. 137–144.
- [154] B. Shuai, A. Bergamo, U. Buechler, A. Berneshawi, A. Boden, and J. Tighe. “Large Scale Real-World Multi-person Tracking”. In: *European Conference on Computer Vision (ECCV)*. 2022, pp. 504–521.
- [155] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski. “DINOv3”. In: *arXiv preprint arXiv:2508.10104* (2025).
- [156] M. Simon, K. Amende, A. Kraus, J. Honer, T. Samann, H. Kaulbersch, S. Milz, and H. Michael Gross. “Complexer-YOLO: Real-Time 3D Object Detection and Tracking on Semantic Point Clouds”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPR)*. 2019.
- [157] K. Simonyan and A. Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [158] G. Singh, S. Akrigg, M. Maio, V. Fontana, R. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, T. Nicholson, J. Omokeowa, S. Grazioso, A. Bradley, G. Gironimo, and F. Cuzzolin. “ROAD: The Road Event Awareness Dataset for Autonomous Driving”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 45:01 (2023), pp. 1036–1054.
- [159] S. Singh. *Critical reasons for crashes investigated in the national motor vehicle crash causation survey*. Tech. rep. U.S. National Highway Traffic Safety Administration (NHTSA), 2015.
- [160] V. Somers, C. De Vleeschouwer, and A. Alahi. “Body Part-Based Representation Learning for Occluded Person Re-Identification”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 1613–1623.
- [161] Z. Song et al. “Robustness-Aware 3D Object Detection in Autonomous Driving: A Review and Outlook”. In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* 25:11 (2024), pp. 15407–15436.

- [162] D. Stadler and J. Beyerer. “BYTEv2: Associating More Detection Boxes Under Occlusion for Improved Multi-person Tracking”. In: *Proc. of the International Conf. on Pattern Recognition (ICPR)*. Lecture Notes in Computer Science. 2023, pp. 79–94.
- [163] O. Styles, V. Sanchez, and T. Guha. “Multiple Object Forecasting: Predicting Future Object Locations in Diverse Environments”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 690–699.
- [164] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. “Scalability in perception for autonomous driving: Waymo open dataset”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 2446–2454.
- [165] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo. “TransTrack: Multiple Object Tracking with Transformer”. In: *arXiv preprint arXiv:2012.15460* (2021).
- [166] J. Suurballe. “Disjoint paths in a network”. In: *Networks* 4:2 (1974), pp. 125–145.
- [167] R. Tao, E. Gavves, and A. W. M. Smeulders. “Siamese Instance Search for Tracking”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2016).
- [168] B. Thérien et al. “Object Re-Identification from Point Clouds”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2024, pp. 8377–8388.
- [169] H. Touvron et al. “Training Data-Efficient Image Transformers & Distillation through Attention”. In: *Proc. of the International Conf. on Machine Learning (ICML)*. 2021, pp. 10347–10357.
- [170] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. “A Siamese Long Short-Term Memory Architecture for Human Re-Identification”. In: *European Conference on Computer Vision (ECCV)* (2016), pp. 135–153. eprint: [1607.08381](#).
- [171] P. M. Visscher. “Sizing up human height variation”. In: *Nature Genetics* (2008).
- [172] C. Vondrick and D. Ramanan. “Video Annotation and Tracking with Active Learning”. In: *Adv. in Neural Information Processing Systems (NeurIPS)*. 2011, pp. 28–36.
- [173] C. Vondrick, D. Ramanan, and D. Patterson. “Efficiently Scaling Up Video Annotation with Crowdsourced Marketplaces”. In: *European Conference on Computer Vision (ECCV)*. 2010, pp. 610–623.
- [174] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, et al. “Constrained k-means clustering with background knowledge”. In: *Proc. of the International Conf. on Machine Learning (ICML)*. Vol. 1. 2001, pp. 577–584.
- [175] G. Wang et al. “Learning discriminative features with multiple granularities for person re-identification”. In: *Proc. of the ACM International Conference on Multimedia*. 2018, pp. 274–282.
- [176] Y.-H. Wang et al. “SMILEtrack: SiMilarity LEarning for Occlusion-Aware Multiple Object Tracking”. In: *Proc. of the AAAI Conference on Artificial Intelligence*. 2024, pp. 5740–5748.

- [177] L. Wang, X. Zhang, W. Qin, X. Li, J. Gao, L. Yang, Z. Li, J. Li, L. Zhu, H. Wang, and H. Liu. "CAMO-MOT: Combined Appearance-Motion Optimization for 3D Multi-Object Tracking With Camera-LiDAR Fusion". In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* (2023), pp. 1–16.
- [178] L. Wang, W. Ouyang, X. Wang, and H. Lu. "Visual Tracking with Fully Convolutional Networks". In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2015, pp. 3119–3127.
- [179] N. Wang, S. Li, A. Gupta, and D.-Y. Yeung. "Transferring Rich Feature Hierarchies for Robust Visual Tracking". In: *arXiv preprint arXiv:1501.04587* (2015).
- [180] N. Wang and D.-Y. Yeung. "Learning a Deep Compact Image Representation for Visual Tracking". In: *Adv. in Neural Information Processing Systems (NeurIPS)*. Vol. 26. 2013.
- [181] P. Wang et al. "Quality-Aware Part Models for Occluded Person Re-Identification". In: *IEEE Transactions on Multimedia* (2023), pp. 3154–3165.
- [182] W. Wang et al. "DMM: Dual-Modal Model for Person Re-Identification". In: *International Joint Conference on Neural Networks (IJCNN)*. 2022, pp. 1–8.
- [183] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu. "DriveDreamer: Towards Real-World-Drive World Models for Autonomous Driving". In: *European Conference on Computer Vision (ECCV)*. Vol. 15106. 2025, pp. 55–72.
- [184] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson. "Ranked list loss for deep metric learning". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5207–5216.
- [185] Y. Wang et al. "TOP-ReID: Multi-Spectral Object Re-identification with Token Permutation". In: *Proc. of the AAAI Conference on Artificial Intelligence* 38:6 (2024), pp. 5758–5766.
- [186] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang. "Towards Real-Time Multi-Object Tracking". In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 107–122.
- [187] N. Wojke, A. Bewley, and D. Paulus. "Simple Online and Realtime Tracking with a Deep Association Metric". In: *Proc. of the International Conf. on Image Processing (ICIP)*. 2017, pp. 3645–3649.
- [188] A. Wu, W.-S. Zheng, and J. Lai. "Robust Depth-based Person Re-identification". In: *IEEE Transactions on Image Processing* 26:6 (2017), pp. 2588–2603.
- [189] X. Wu and J. Xu. "P-IoU: Accurate Motion Prediction Based Data Association for Multi-object Tracking". In: *International Conference on Neural Information Processing (ICONIP)*. 2024, pp. 484–496.
- [190] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao. "DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model". In: *IEEE Robotics and Automation Letters (RA-L)* 9:10 (2024), pp. 8186–8193.

- [191] Y. Xue, G. Luo, Q. Yuan, J. Li, and F. Yang. “ST-Tracking: Spatial-temporal Graph Convolution Neural Network for Multi-object Tracking”. In: *Proc. of the IEEE Intelligent Transportation Systems Conf. (ITSC)*. 2021, pp. 2958–2964.
- [192] 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”. In: *arXiv preprint arXiv:2305.12724* (2023).
- [193] B. Yang, M. Bai, M. Liang, W. Zeng, and R. Urtasun. “Auto4D: Learning to Label 4D Objects from Sequential Point Clouds”. In: *arXiv preprint arXiv:2101.06586* (2021).
- [194] B. Yang, J. Chen, and M. Ye. “Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 16870–16879.
- [195] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng. “Track Anything: Segment Anything Meets Videos”. In: *arXiv preprint arXiv:2304.11968* (2023).
- [196] Q. Yang et al. “Focus on the Visible Regions: Semantic-Guided Alignment Model for Occluded Person Re-Identification”. In: *Sensors* (2020), p. 4431.
- [197] T. Yau, S. Malekmohammadi, A. Rasouli, P. Lakner, M. Rohani, and J. Luo. “Graphsim: A graph-based spatiotemporal interaction modelling for pedestrian action prediction”. In: *Proc. of the International Conf. on Robotics and Automation (ICRA)*. 2021, pp. 8580–8586.
- [198] M. Ye et al. “Deep learning for person re-identification: A survey and outlook”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 44:6 (2021), pp. 2872–2893.
- [199] S. Yi, H. Li, and X. Wang. “Pedestrian Behavior Understanding and Prediction with Deep Neural Networks”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 263–279.
- [200] 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: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 2636–2645.
- [201] L. Yue, M. Abdel-Aty, Y. Wu, O. Zheng, and J. Yuan. “In-Depth Approach for Identifying Crash Causation Patterns and Its Implications for Pedestrian Crash Prevention”. In: *Journal of Safety Research* 73 (2020), pp. 119–132.
- [202] J. Yuen, B. Russell, and A. Torralba. “LabelMe video: Building a video database with human annotations”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2009, pp. 1451–1458.
- [203] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei. “MOTR: End-to-End Multiple-Object Tracking with Transformer”. In: *European Conference on Computer Vision (ECCV)*. 2022, pp. 659–675.
- [204] C. Zhang, H. Wang, L. Chen, Y. Li, and Y. Cai. “MixedFusion: An Efficient Multimodal Data Fusion Framework for 3-D Object Detection and Tracking”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2023), pp. 1–15.

- [205] L. Zhang, Y. Li, and R. Nevatia. “Global data association for multi-object tracking using network flows”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2008, pp. 1–8.
- [206] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu. “Magic Tokens: Select Diverse Tokens for Multi-modal Object Re-Identification”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 17117–17126.
- [207] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. C. Loy. “Robust Multi-Modality Multi-Object Tracking”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2019, pp. 2365–2374.
- [208] X. Zhang, Z. Fan, X. Tan, Q. Liu, and Y. Shi. “Spatiotemporal Adaptive Attention 3D Multiobject Tracking for Autonomous Driving”. In: *Knowledge-Based Systems* 267 (2023), p. 110442.
- [209] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. “ByteTrack: Multi-Object Tracking by Associating Every Detection Box”. In: *European Conference on Computer Vision (ECCV)*. 2022, pp. 1–21.
- [210] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu. “FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking”. In: *International Journal of Computer Vision* 129:11 (2021), pp. 3069–3087.
- [211] Y. Zhang, X. Wang, X. Ye, W. Zhang, J. Lu, X. Tan, E. Ding, P. Sun, and J. Wang. “ByteTrackV2: 2D and 3D Multi-Object Tracking by Associating Every Detection Box”. In: *arXiv preprint arXiv:2303.15334* (2023).
- [212] S. Zhao, Y. Wu, S. Wang, W. Ke, and H. Sheng. “Mask Guided Spatial-Temporal Fusion Network for Multiple Object Tracking”. In: *Proc. of the International Conf. on Image Processing (ICIP)*. 2022, pp. 3231–3235.
- [213] L. Zheng et al. “Scalable Person Re-identification: A Benchmark”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2015, pp. 1116–1124.
- [214] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. “MARS: A Video Benchmark for Large-Scale Person Re-identification”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 868–884.
- [215] Z. Zheng, L. Zheng, and Y. Yang. “A Discriminatively Learned CNN Embedding for Person Reidentification”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2018), pp. 1–20.
- [216] S. Zhou et al. “Depth Occlusion Perception Feature Analysis for Person Re-Identification”. In: *Pattern Recognition Letters* (2020), pp. 617–623.
- [217] X. Zhou, V. Koltun, and P. Krähenbühl. “Tracking Objects as Points”. In: *European Conference on Computer Vision (ECCV)* (2020), pp. 474–490.
- [218] H. Zhu et al. “SEAS: ShapE-Aligned Supervision for Person Re-Identification”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 164–174.
- [219] K. Zhu et al. “AAformer: Auto-Aligned Transformer for Person Re-Identification”. In: *IEEE Transactions on Neural Networks and Learning Systems* 35:12 (2023), pp. 17307–17317.

- [220] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. “Deformable {DETR}: Deformable Transformers for End-to-End Object Detection”. In: *International Conference on Learning Representations (ICLR)*. 2021.
- [221] Y. Zhu, Q. Wang, C. Li, J. Tang, C. Gu, and Z. Huang. “Visible-thermal multiple object tracking: Large-scale video dataset and progressive fusion approach”. In: *Pattern Recognition* 161 (2025), p. 111330.
- [222] Z. Zhu, J. Nie, H. Wu, Z. He, and M. Gao. “MSA-MOT: Multi-Stage Association for 3D Multimodality Multi-Object Tracking”. In: *Sensors* 22:22 (2022), p. 8650.

