UNIVERSITY OF SALERNO

DEPARTMENT OF INFORMATION AND ELECTRICAL ENGINEERING AND APPLIED MATHEMATICS

MASTER'S DEGREE COURSE IN COMPUTER ENGINEERING

MASTER'S THESIS

ON

AUTONOMOUS VEHICLES DRIVING

Semi-Automatic labeling of dataset acquired by ego-vehicle via transfer learning

Supervisors:

Prof. Gragnaniello

Prof. Vento

Candidate:

Sica Ferdinando

Matr. 0622701794

Academic Year 2022/2023

"In the end, it's not the years in your life that count. It's the life in your years."

- Abraham Lincoln

Abstract

Description of the problem faced

In the technological revolution of the 21st century, autonomous driving represents one of the central pillars that is redefining the entire transport sector, with repercussions that go far beyond, influencing the way we design our cities, infrastructure, and our daily lifestyle. The promise of vehicles that can autonomously navigate through complex environments is captivating. However, behind this promise lies a major challenge: the significance and complexity of data. One of the main players in this revolution, often hidden but of vital importance, is the world of data. Data powers every aspect of autonomous driving, from recognizing obstacles and pedestrians, to route planning, to the real-time decisions a vehicle must make in unpredictable situations.

Framing of the paper in the contemporary technical scenario

For data to serve its purpose effectively, it needs to be precise, pertinent, and properly annotated. The task of annotating data presents multiple complexities. Given the surge in data availability, the manual annotation of each data point soon becomes unfeasible. Mistakes due to human fatigue, repetitive nature of the task, and oversights can result in errors that, especially in scenarios like autonomous driving, can lead to severe repercussions.

Personal contribution of the candidate to the solution of the problem described

Faced with this challenge, our research sought to devise a solution that could balance efficiency with accuracy. Through the adoption of transfer learning and the integration of complex algorithms, we aimed to address potential network issues and

bring artificial intelligence to the forefront of the labeling process. Our objectives were twofold: firstly, to diminish the manual workload for annotators, and secondly, to refine the quality of the associated data.

Description of the application/experimental content of the paper

The experiments conducted and described within this thesis demonstrate the feasibility and effectiveness of this approach. We saw how, in certain scenarios, our semi-automatic labeling framework could not only speed up the labeling process but also improve the overall accuracy of the labels. Beyond tangible results, this thesis also represents a broader reflection on the role of data in the age of autonomous driving. With the advent of increasingly intelligent and connected vehicles, the demand for accurate and well-labeled data will grow exponentially. The need for efficient labeling methods will therefore become increasingly pressing.

The research presented offers an in-depth look at the challenges and opportunities in data labeling for autonomous driving. It marks a significant step towards creating more effective and efficient methods, laying the groundwork for future developments in this crucial field.

Contents

Chapter 1

Introduction

The field of computer engineering has seen explosive growth in the last decade, with increasingly broad and sophisticated applications of artificial intelligence (AI). One of the most fascinating and promising areas in this context is autonomous driving, aiming to revolutionize the automotive industry and mobility in general. However, at the heart of every autonomous driving system lies data, as it's through these data that AI learns to drive. Specifically, the quality and accuracy of the labels attributed to the data are crucial for training reliable and safe artificial intelligence models.

This thesis sets the ambitious goal of addressing one of the most significant challenges in autonomous driving: dataset labeling. This issue is crucial, as the labeling process requires significant human effort and resources. The aim of this thesis is to develop a semi-automatic labeling system that leverages the potential of modern neural networks and machine learning to automate much of the labeling process, thereby reducing the time and resources required.

Accurate and comprehensive data labeling is essential to ensure the safety and effectiveness of autonomous vehicles. These vehicles must be able to recognize and respond to a wide range of objects and situations on the road, including pedestrians, vehicles, traffic lights, road signs, and much more. For autonomous driving models to learn to correctly recognize and interpret these situations, an extremely rich and detailed training dataset is required. However, manually labeling such a dataset would require a prohibitive amount of time and resources.

In this context, the proposed approach presents itself as an innovative and

effective solution. The system will consider using models pre-trained on specific datasets, like BDD100K [42], to generate initial labels for the reference dataset, specifically the Waymo dataset. This step is crucial as it utilizes the knowledge transfer from pre-trained models, allowing for high-quality initial labels. As we will see in the dedicated chapters, the main difference between the two mentioned datasets lies in the difference in "classes"; specifically, the Waymo dataset has fewer labeled "classes" compared to BDD100K. For instance, Waymo recognizes "vehicles" while BDD100K differentiates "car", "truck", and "bus."

However, the most innovative and exciting aspect of this thesis is the implementation of rule-based algorithms to automatically optimize and correct the initial labels generated by the pre-trained models. This phase is a milestone in automating the labeling process, as it aims to minimize human intervention. The goal is to ensure that the labels are not only accurate but also consistent and detailed.

Throughout the document, we will address the main problems related to semi-automatic labeling and then delve into the specific problems we tried to solve related to the particular model used. We will analyze the state of the art for Multi-Object Tracking (MOT) and the BDD100K dataset, the state of the art for automatic labeling, the state of the art for Video Annotation Tools (VAT), and much more.

1.1 Problem Definition

The fundamental problem this thesis aims to address is the process of semi-automatic data labeling, with a specific application to autonomous driving but with broader implications for any context that requires the annotation of complex datasets. Data labeling is a critical phase in preparing datasets used to train artificial intelligence models, but it's often a costly, error-prone, and time-intensive process. The problem can be broken down into several key challenges:

High Cost and Resource Expenditure in Manual Data Labeling The high cost and resource expenditure associated with manual data labeling represent one of the most significant challenges in contexts where acquiring labeled datasets is crucial. This issue is particularly evident in autonomous driving, where it's essential to collect and annotate vast amounts of data from various sources, such as cameras,

LiDAR sensors, and radar sensors (the latter not addressed in this document), to train autonomous vehicles to operate in a wide range of real-world scenarios.

1. **Human Resources:** A key factor contributing to the high cost is the need for highly qualified personnel to carry out manual data labeling. Annotators must be trained to understand the details of the specific context and the ontology of the required labels, which takes considerable time and investment from the companies and institutions involved. Furthermore, the availability of qualified annotators can be limited, leading to further delays in the labeling process.

2. **Time:** Manual labeling is time-consuming. The need to carefully examine each sample and apply the correct labels can significantly slow the progress of the autonomous driving project. Additionally, research and development teams must coordinate and manage human resources to ensure datasets are annotated promptly and consistently.

3. **Financial Costs:** The financial investments required to cover the costs of human resources involved in manual data labeling can be considerable. These costs include wages, training, IT infrastructure, and administrative expenses. For companies operating on a limited budget, this expenditure can pose a significant obstacle.

4. **Limited Scalability:** The scalability of manual labeling is limited by the availability of qualified human resources. Increasing the number of annotators can lead to challenges in managing and quality control of labels, and may even increase the risk of errors due to inconsistencies between annotators.

From a scientific perspective, the challenge of manual data labeling translates into a significant obstacle for research and development in the field of autonomous driving. Science demands continuous access to high-quality and varied datasets to advance understanding of problems and to develop innovative solutions. Manual labeling can limit the quantity and speed at which researchers can access critical data for the development of advanced autonomous driving algorithms.

Addressing this challenge is essential for the scientific evolution of the field of autonomous driving and other areas where manual data labeling represents a

bottleneck. The solution proposed in this thesis, which is based on reducing the dependency on human annotators through the implementation of a semi-automatic system, aims to support scientific research by allowing more efficient data collection and faster access to high-quality labeled datasets.

Susceptibility to Human Errors The susceptibility to human errors is a critical issue in manual data labeling, which has severe implications in the context of autonomous driving and other sectors where accurately labeled data are essential for the training and operation of advanced artificial intelligence models.

1. **Labeling Errors:** Humans, even if highly qualified, can make mistakes during the data labeling process. These errors can stem from various sources, such as fatigue, distraction, or misunderstanding of the specific labels to be applied. Such mistakes can adversely affect data quality and result in unsatisfactory performance in autonomous driving models.

2. **Inconsistency among Annotators:** When multiple annotators work on a dataset, inconsistencies in the labels applied are common. These inconsistencies can arise from differences in interpreting annotation guidelines or discrepancies in individual perceptions of scenarios. Inconsistency among annotators can make datasets less reliable for scientific research and algorithm development.

3. **Difficulty in Quality Control:** Quality control in manual data labeling can be complex and requires additional resources. Systematic verification of label accuracy can demand significant time and effort, leading to further delays in data availability for scientific research and development purposes.

From a scientific standpoint, susceptibility to human errors poses a significant challenge in conducting experiments and studies in the field of autonomous driving and machine learning. The presence of errors in training data can lead to skewed results and misleading conclusions in scientific research. Additionally, inconsistency among annotators can make replicating results and sharing datasets within the scientific community challenging.

The solution proposed in this thesis, aiming to assist the human through automation and to introduce a semi-automatic systems for data labeling, has the potential to

significantly enhance the quality of datasets used in scientific research. This would enable researchers to conduct more reliable experiments and obtain more robust scientific findings in the realm of autonomous driving and other related disciplines.

Extended Model Development Cycle The extended model development cycle poses a significant challenge in the context of semi-automatic data labeling for applications such as autonomous driving. From a scientific perspective, this challenge impacts several aspects of research and development in the fields of artificial intelligence and machine learning.

1. **Temporal and Financial Resources:** Developing autonomous driving models demands substantial resources in terms of time and money. The extended development cycle, which can span months or even years, can significantly slow down scientific research and project iteration. The need to wait for the cycle's completion to evaluate model performance can be a hindrance to rapid iteration and enhancement.

2. **Risk of Obsolescence:** In the rapidly evolving field of autonomous driving, models and algorithms must be continually updated to remain competitive and address new challenges. The extended model development cycle risks rendering the models obsolete or less effective by the time they are completed.

3. **Experimentation Challenges:** Scientists and researchers might wish to conduct quick experiments to test new ideas or training strategies. However, the extended development cycle can impede such flexibility, making it difficult to swiftly conduct tests and scientific evaluations.

From a scientific standpoint, the extended model development cycle presents challenges for the validity of experiments and scientific conclusions. The delay in obtaining results and evaluating new ideas can limit the ability to conduct meaningful experiments and respond timely to discoveries.

The proposed approach of semi-automatic data labeling aims to mitigate these scientific challenges by reducing the model development cycle. Automating data labeling and employing semi-automatic systems can accelerate the collection of training data and the evaluation of model performance. This offers researchers the

opportunity to conduct faster experiments, while simultaneously reducing the risk of model obsolescence. Such an approach could promote more agile and dynamic scientific research in the realms of autonomous driving and artificial intelligence.

Need for High-Quality Datasets The availability of high-quality datasets plays a crucial role in the context of semi-automatic data labeling for applications such as autonomous driving. From a scientific perspective, the quality of the data used to train and evaluate models significantly impacts the validity of experiments and the reliability of discoveries. This section examines the scientific aspect related to the need for high-quality datasets in this context.

1. **Reliability of Evaluations:** In autonomous driving contexts, safety is of paramount importance. Collecting high-quality data is essential for reliably assessing the performance of autonomous driving models. The presence of erroneous data or inaccurately assigned labels can lead to misleading evaluations and jeopardize road safety. This scientific aspect underscores the importance of dataset quality in autonomous driving research.

2. **Generation of Realistic Data:** Datasets used to train autonomous driving models must accurately represent real road conditions. Lack of realism in training data can result in models that are not robust and cannot handle complex and unexpected situations. From a scientific standpoint, creating high-quality datasets that faithfully capture the variability of the real world is a significant challenge.

3. **Model Generalization:** Scientific research often aims to develop autonomous driving models capable of effectively generalizing across a wide range of road situations. The quality of the datasets used for training plays a crucial role in a model's ability to generalize. Low-quality or limited datasets can lead to models that struggle with generalization, limiting the scientific applicability of the results obtained.

The need for high-quality datasets is a relevant scientific challenge in the field of autonomous driving. The proposed semi-automatic data labeling approach seeks to address this challenge by improving the quality and accuracy of data labeling,

ensuring that the datasets used for training are reliable, realistic, and promote model generalization. In this way, the goal is to enhance the scientific validity of research and contribute to the advancement of autonomous driving technologies.

To tackle this challenge, the solution proposed in this thesis relies on reducing dependence on human annotators through the implementation of a semi-automatic system that leverages artificial intelligence to generate initial labels. This approach aims to mitigate the costs and intensive use of human resources, allowing teams to focus on reviewing and optimizing automatically generated labels. Thus, it seeks to address the high cost and resource-intensive nature associated with manual data labeling, paving the way for greater efficiency and scalability in acquiring high-quality datasets.

1.2 Relevance of the Issue in Computer Engineering and Autonomous Driving

The field of computer engineering has witnessed an explosion of innovations in recent decades, with Autonomous Driving emerging as one of the most promising and revolutionary sectors. This technology aims to radically transform the way we conceive road transportation, making it safer, more efficient, and sustainable. However, like any technological revolution, Autonomous Driving is accompanied by difficult challenges, including the need for high-quality datasets for model training.

1.2.1 Relevance in Computer Engineering

In the field of Computer Engineering, Autonomous Driving represents one of the most significant challenges and opportunities. This revolutionary technology combines concepts of real-time data processing, deep neural networks, and machine learning algorithms to create systems capable of making complex decisions in a dynamic context such as driving on public roads. Autonomous Driving relies on sophisticated sensors such as cameras, LiDAR, radar, and high-precision GPS positioning systems to perceive the surrounding environment and drive the vehicle autonomously. In this context, high-quality datasets play a crucial and multi-dimensional role.

Development of Computer Vision Algorithms Autonomous Driving requires accurate identification and interpretation of road signs, vehicles, pedestrians, and obstacles. Computer vision algorithms are essential for the recognition and understanding of this information. Computer engineering has made significant advances in the field of computer vision with the advent of deep learning. High-quality datasets containing a wide variety of road situations, weather conditions, and moving objects are essential to train deep neural networks capable of reliably recognizing elements of the road context.

Control and Decision Systems Autonomous vehicles must make real-time decisions based on information perceived from the surrounding environment. These decisions involve speed, direction, overtaking, braking, and more. Autonomous vehicle control algorithms rely on accurate data to determine how to respond to complex road situations. High-quality datasets allow the training of decision models that take into account a wide range of variables, contributing to improving the safety and efficiency of the system.

Testing and Simulations The testing phase is essential for fine-tuning Autonomous Driving systems. However, on-road experimentation presents safety and logistical challenges. Computer engineering addresses this issue by developing advanced simulations based on real data. Accurate and detailed datasets are necessary to create realistic simulations in which autonomous vehicles can be tested in virtual conditions covering a wide range of scenarios. These tests allow engineers to identify and resolve potential issues safely and efficiently before implementing solutions in the real world.

Technological Advancements Computer engineering is intrinsically linked to technological innovation. Solving the challenge of dataset quality in Autonomous Driving paves the way for significant technological developments. From improving the perception of autonomous vehicles to creating more sophisticated control algorithms, computer engineering crucially contributes to the advancement of Autonomous Driving technology. High-quality datasets represent the fuel that drives this innovation, enabling researchers to push beyond the limits of current performance.

1.2.2 Relevance in Autonomous Driving

Autonomous Driving represents one of the greatest innovations in the automotive and mobility sector. This technology is radically transforming our conception of road transportation, offering promises of increased safety, efficiency, and accessibility. To fully understand the importance of high-quality datasets in Autonomous Driving, it is essential to explore the unique challenges this technology faces and the motivations that make it a priority.

Road Safety Autonomous Driving has been primarily conceived to improve road safety. Road accidents are one of the leading causes of death worldwide, and many of these tragedies are caused by human errors. Autonomous vehicles promise to drastically reduce human involvement in driving decisions, but to do so reliably, they require high-quality data. The datasets used for training autonomous driving models must accurately represent a wide range of road situations, weather conditions, and driver behaviors. Only then can autonomous vehicles learn to react safely and effectively, thus helping to reduce the number of road accidents.

Understanding the Road Context One of the primary challenges for autonomous vehicles is the accurate understanding of the road context. This involves recognizing road signs, understanding the intentions of other drivers, detecting obstacles, and interpreting the surrounding terrain. Addressing these challenges requires access to high-quality datasets that capture all these dynamics in detail. Without accurate data, autonomous vehicles may struggle to distinguish an emergency situation from a normal braking maneuver or to recognize the difference between a green and a red traffic light. The quality of the datasets is therefore crucial to ensure a reliable understanding of the road context.

Adaptation to Variable Conditions Road conditions can vary widely, from a sunny day to a snowstorm, from an empty road to a crowded intersection. Autonomous Driving must be able to adapt to these conditions smoothly and safely. High-quality datasets capture this variability, allowing autonomous vehicles to be trained on a wide range of situations. This is essential because an Autonomous Driving system that works well in one situation may fail in another. Adaptability is one of the key

advantages of the technology, and data plays a central role in this learning process.

Regulatory Compliance The widespread adoption of Autonomous Driving requires compliance with strict regulations and standards. Regulatory authorities often require reliable evidence of the effectiveness and safety of autonomous vehicles before allowing them to operate on public roads. High-quality datasets are an essential tool for demonstrating the compliance and safety of autonomous vehicles. They can be used to conduct comprehensive test simulations, demonstrate the vehicles' ability to handle critical scenarios, and ultimately obtain approval from the relevant authorities.

In conclusion, the relevance of high-quality datasets in Autonomous Driving is undeniable. These data serve as the foundation for the success and safe adoption of this revolutionary technology. The unique challenges faced by Autonomous Driving require a solid data foundation to build upon, and only then can we fully realize the potential of safer, more efficient, and sustainable driving.

Chapter 2

State of art

In the context of computer engineering and autonomous driving, the critical importance of data, particularly labeled data, cannot be underestimated. In recent years, both the industry and the scientific community have recognized the need to expedite the data labeling process or even minimize human involvement to create increasingly extensive and precise datasets. This recognition has led to a wide range of scientific publications and the development of dedicated tools aimed at addressing this challenge. However, despite the progress made, none of these existing approaches fully meet the specific requirements we set out to address in the course of this research.

Over the years, the primary goal has been to simplify the work of annotators by attempting to develop tools that can assist them in labeling complete datasets. Some of these tools enable semi-automatic labeling, a process in which an artificial intelligence model performs initial data labeling, while annotators intervene to improve and correct it. This approach represents a significant time and human resource-saving method. Some tools, such as those mentioned in [17] and [39], provide this functionality, allowing the use of pre-existing models for the initial labeling of data, followed by a human review phase.

One of the most recent and innovative tools in this field is [11], as described in CVAT: Computer Vision Annotation Tool. CVAT stands out for its intuitive graphical interface and support for a wide range of annotation types. This tool also allows users to incorporate custom models (though in limited numbers) for semi-automatic labeling or to import existing annotations directly. However, once again, a fundamental component is missing: techniques for the automatic improvement

of uploaded or generated annotations.

A tool that comes closer to our vision is an implementation of [39] described in [40]. This tool offers advanced algorithms to analyze available data and suggest possible errors made by the model, leaving the responsibility of correcting them to the annotator. However, the process remains largely manual, and what we aim to achieve with this research is the automation of this critical phase.

During the literature review, we also focused on the models used for both the Detection and Multi-Object Tracking (MOT) phases. Detection is a fundamental step for identifying objects in images or videos. Over the years, various approaches have been developed, including Convolutional Neural Networks (CNNs) [16], Recurrent Neural Networks (RNNs), and Transformer-based networks [38]. Additionally, "Two Stage Detectors" and "Single Stage Detectors" have been introduced to address this task. The latter, in particular, have framed it as a regression problem and led to significant innovations such as "Focal Loss" to tackle class imbalance.

Further progress has been driven by Transformers, originally developed for Natural Language Processing (NLP). These models have been successfully adapted for Detection tasks, addressing some of the limitations of CNN-based networks.

Following the Detection phase, the challenge of Multi-Object Tracking (MOT) begins, aiming to maintain the consistency of object trajectories over time, despite difficulties such as occlusion and environmental variations. In this context, QDTrack [15] represents the most advanced model with outstanding performance on datasets like BDD100K.

This chapter will further explore the state of the art in the field of autonomous driving, focusing on innovative tools and approaches for data management and the evolution of models used for Detection and MOT. The aim is to provide a comprehensive overview of existing solutions and the latest advancements that have led us to develop a unique solution tailored to our specific needs.

2.1 Detailed Analysis of the State of the Art

In this chapter, we will conduct a detailed analysis of the state of the art in three key areas: Detection, Multi-Object Tracking, and Video Annotation Tools. We will

explore the most relevant methods, algorithms, and tools in each of these categories, highlighting the advancements and challenges faced in the fields of computer engineering and autonomous driving.

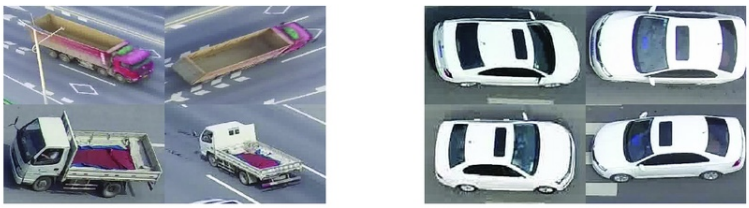
2.2 Detection

Detection is a critical phase in the perception of the environment by autonomous vehicles. This section will examine the most significant developments in the field of Detection, with a particular focus on the models and techniques used to detect and locate objects in images and videos. Object detection is a natural extension of object classification, which aims only to recognize the object in the image. The goal of object detection is to identify all instances of predefined classes and provide their approximate location in the image through axis-aligned boxes (commonly known as Bounding Boxes). The detector should be capable of identifying all instances of object classes and draw a bounding box around them. In general, this problem is viewed as a supervised learning problem. Modern object detection models have access to large sets of labeled images for training and are evaluated on various canonical benchmarks.

2.2.1 Key Challenges in Object Detection

Computer vision has made significant progress in the past decade, but it still faces some key challenges in real-world applications. Some of these fundamental challenges encountered by networks in real-world applications include:

- **Intra-Class Variability:** Intra-class variability among instances of the same object is relatively common in nature. This variability can be due to various reasons such as occlusion, lighting, pose, viewpoint, etc. These uncontrolled external factors can have a dramatic effect on the appearance of the object. Objects may undergo non-rigid deformations or be rotated, resized, or blurred. Some objects may have unclear boundaries, making extraction challenging. Here are some examples 2.1:



(a) Figure from [33] Intra-class variation and inter-class similarity in vehicle identification.

(b) Figure from [21] Example of intra-class variation.

(c) Figure from [25] Intra-class variation in chair classification.

Figure 2.1: Illustrations of intra-class variation in different contexts.

- **Number of Categories:** The number of object classes available for classification makes the problem a challenging one to tackle. Furthermore, it requires a

larger amount of high-quality annotated data, which can be difficult to obtain. The use of a limited number of examples to train a detector remains an open research issue.

- **Efficiency:** Current models require significant computational resources to generate accurate detection results. With mobile and edge devices becoming increasingly prevalent, efficient detectors are crucial for the future development of computer vision.

2.2.2 Metrics

Object detectors use multiple criteria to measure performance, including frames per second (FPS), precision, and recall. However, mean Average Precision (mAP) is the most common evaluation metric.

Precision (2.1) is defined as the ratio of True Positives to the sum of True Positives and False Positives. In other words, it measures the percentage of correct predictions made by the detector. Higher precision indicates the detector's ability to avoid false alarms.

Recall (2.2), also known as Sensitivity, is defined as the ratio of True Positives to the sum of True Positives and False Negatives. It represents the detector's ability to correctly detect all objects in the images. A high recall indicates a lower likelihood of missing objects during detection.

The mean Average Precision (mAP) is a composite measure that takes into account precision at various IoU thresholds. Intersection over Union (IoU) is a parameter that determines how much two bounding boxes overlap and is used to determine how much the predicted bounding box must overlap with the ground truth to be considered a correct detection. mAP is calculated as the average of precisions calculated at different IoU thresholds. This means that a detector must be accurate at different overlap scales to achieve a high mAP score. mAP is an important metric for evaluating object detection performance.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2.1}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2.2}$$

2.2.3 Object Detectors

In this section, we will examine the landscape of object detectors, initially focusing on pioneering works that laid the foundation for subsequent developments. These detectors have played a crucial role in the history of object detection, contributing to defining fundamental concepts and techniques used in modern models.

Object detection is an essential component in the field of computer vision, with applications ranging from security to autonomous driving. Before exploring the advanced techniques used in modern detectors, it is important to understand the evolution of this field and how early works helped define the initial challenges and solutions.

1. **Viola-Jones:** Proposed in 2001, the Viola-Jones object detector [31] was one of the first accurate and powerful detectors. Although primarily designed for face detection, this algorithm played a fundamental role in the history of object detection. Using Haar-like features, integral images, Adaboost, and a cascaded classifier, the Viola-Jones detector was able to achieve high performance efficiently. This pioneering work is still cited and used as a reference in areas where efficiency is crucial.

2. **HOG Detector:** In 2005, Dalal and Triggs introduced Histogram of Oriented Gradients (HOG) [12] as a feature descriptor for object detection. While the HOG detector is primarily known for pedestrian detection, it represented a significant step toward object detection in general. This method demonstrated the importance of extracting discriminative features to improve detection accuracy.

3. **DPM:** The Deformable Part Model (DPM) [14], introduced by Felzenszwalb et al., represented an innovative approach to object detection. Based on the divide-and-conquer philosophy, DPM showed that detection could be improved by decomposing objects into recognizable parts and combining the information to achieve more accurate detection. This approach achieved remarkable results and influenced the future development of object detectors.

These three pioneering works played a crucial role in the early development of object detection and established important principles and techniques used in modern detectors. Now, we will examine some of the more advanced techniques and innovations that marked the transition to more sophisticated and accurate models.

2.2.4 Convolutional Neural Networks (CNN)

Convolutional Neural Networks, commonly referred to as CNNs, have revolutionized the field of object detection, enabling greater accuracy and efficiency in identifying objects in images and videos. This section will focus on the pivotal role played by CNNs in detection applications, exploring their key features and the backbone architectures used for feature extraction.

CNNs draw inspiration from biology and human visual perception. They are designed to automatically learn discriminative features directly from images, reducing the need for manual feature extraction. Learned features include edges, textures, and shapes, which are crucial for object detection.

One of the key aspects of CNNs is the presence of convolutional layers that apply convolutional filters to input images. These filters enable the capture of local patterns and features, such as lines, curves, and textures. Convolutional layers are followed by pooling layers, which progressively reduce the size of feature maps while preserving the most relevant features.

Backbone architectures are one of the most important elements of object detectors, as they extract features from the input image used by the model. These architectures play a crucial role in the overall effectiveness of the detector. Some of the most significant backbone architectures used in modern detectors include:

- **ResNet:** Residual Network (ResNet) is one of the most influential and widely used architectures. It introduces the concept of "skip connection", allowing direct information flow between layers. This helps mitigate the vanishing gradient problem and enables the training of very deep networks.

- **VGGNet:** The VGG network is known for its simplicity and is primarily composed of 3x3 convolutional layers with max pooling. It has demonstrated excellent performance in many computer vision tasks.

- **InceptionNet:** Inception networks, also known as GoogleNet, use filters of different sizes in parallel to capture features at various scales. This approach allows for more comprehensive feature extraction from images.

Convolutional Neural Networks (CNNs) have proven to be highly effective in object detection across a variety of contexts, including object recognition, face detection, vehicle identification, and more. Their ability to learn directly from images makes them a powerful tool for object detection.

In the context of autonomous driving, CNNs play a crucial role in the perception of the surrounding environment by autonomous vehicles.They can identify pedestrians, vehicles, road signs, and other objects critical for road safety.

In conclusion, Convolutional Neural Networks represent a milestone in object detection, offering a combination of precision and efficiency. Their central role in the visual perception of autonomous machines makes them a fundamental component in the field of autonomous driving.

2.2.5 Single Stage vs. Two Stage Detectors

In the field of object detection, two main approaches have emerged: "Single Stage Detectors" and "Two Stage Detectors." These approaches represent different design philosophies and have distinctive characteristics that make them suitable for different situations.

Single Stage Detectors

"Single Stage Detectors" are known for their efficiency and processing speed. These models perform object detection in a single pass through the neural network, without the need for a candidate proposal phase. This makes them ideal for real-time applications where speed is crucial, such as autonomous driving.

One of the most well-known "Single Stage Detectors" is YOLO (You Only Look Once) [1], which divides the image into a grid and predicts bounding boxes and object classes directly from each grid cell. This approach is known for its speed and has been widely used in real-time detection applications.

Two Stage Detectors

"Two Stage Detectors" follow a two-phase approach for object detection. In the first phase, these models propose a set of candidates or regions of interest (Region Proposal) that may contain objects. In the second phase, they perform classification and refinement of bounding boxes to identify the objects.

A well-known example of a "Two Stage Detector" is Faster R-CNN [34], which uses a convolutional network for the candidate proposal phase and a second network for classification and refinement of bounding boxes. This approach is known for its high accuracy and is often used in applications where precision is crucial.

2.2.6 Integration of Transformers

The integration of Transformer-based models has represented a significant advancement in object detection, bringing notable improvements in understanding the global context of images. In this section, we will explore how Transformer-based models have been successfully adapted to address challenges and enhance performance in detection tasks.

The Role of Transformers in Detection

Transformer-based models [38], originally known for their excellence in natural language processing tasks, have proven to be highly adaptable for computer vision. The key feature of Transformers is their ability to handle relationships between different parts of an input without relying on a sequential structure. This makes them ideal for tackling the challenges presented by detection tasks, leveraging information from previous inputs to make predictions about the current input.

Transformer in Action

The integration of Transformer-based models in detection has involved the use of hybrid architectures that combine the advantages of Convolutional Neural Networks (CNNs) and Transformers. In these architectures, CNNs are still employed for local feature extraction from the image, while Transformers work on the global representation, capturing context among objects.

A well-known example of a model that integrates Transformers in detection is DETR (DEtection TRansformers). DETR [9] employs a Transformer to directly produce the output of detection, eliminating the need for a candidate proposal step. This approach has demonstrated competitive performance with traditional Two-Stage Detectors.

Advantages of Transformer Integration

The integration of Transformers in object detection has led to several advantages:

- **Improved Context Understanding:** Transformers enable a global understanding of images, helping capture relationships and contexts among objects, which can lead to more accurate detections.
- **Reduced Dependency on Candidate Proposals:** Some Transformer-based models, such as DETR, eliminate the need for a candidate proposal phase, simplifying the detection process.
- **Adaptability:** Transformers are highly adaptable and can be used in combination with other architectures, allowing flexibility in designing custom models for specific detection applications.

Challenges and Future Developments

Despite the evident advantages, the integration of Transformers in object detection still presents some challenges, such as managing computational efficiency and the need for adequate training datasets.

The future of Transformer integration in object detection is promising, with further developments expected to enhance model performance. Ongoing innovation in the field of Transformer-based models opens up new opportunities for advanced detection applications.

2.3 Introduction to Multi-Object Tracking (MOT)

Multi-Object Tracking (MOT) represents one of the fundamental challenges in the field of computer vision. This problem holds great significance in numerous practical

applications, including security monitoring, autonomous driving, video analysis, and more. In general, MOT aims to simultaneously identify, track, and monitor multiple objects within sequences of images or videos.

Contemporary MOT methods primarily follow the "tracking-by-detection" paradigm, which means they detect objects in each frame and subsequently associate them based on the similarity of estimated instances. However, this position-based heuristic works better in simple scenarios and can easily generate errors in the presence of occlusion or crowded scenes.

To address this issue, some methods have introduced motion estimates or path adjustments to ensure accurate distance estimates between objects. However, the aspect similarity aspect often takes a back seat. Search regions are limited to local neighborhoods to avoid distractions since appearance features are not effective at distinguishing different objects. In contrast, humans can associate identical objects based solely on appearance. This is presumed to be due to the suboptimal use of image and object information to learn object similarity.

This traditional approach views instance similarity as a subsequent phase to object detection or uses only scattered ground truth bounding boxes as training samples. These processes ignore the majority of proposed regions in the images, missing out on the opportunity for more effective learning of object similarity.

2.3.1 State of the Art

In the context of Multi-Object Tracking, it is crucial to understand existing methods and the challenges faced by the research community. Some notable approaches include the tracking-by-detection paradigm; however, this approach tends to struggle in complex scenarios with occlusion and crowding.

Some methods have attempted to address these challenges by introducing motion estimates or path adjustments to improve the accuracy of object distances. However, the aspect similarity of objects has often been overlooked or treated only as a support for object association or re-identification of disappeared objects.

In recent research, there has emerged a need to use a wide range of potential regions in frames for supervision in similarity learning, including both positive and negative training samples. This has led to the development of approaches such as

"quasi-dense similarity learning", which relies on associating hundreds of regions of interest between pairs of images for contrastive learning. This method provides broader coverage of informative regions in images and more effective supervision in instance similarity learning.

Furthermore, contrastive learning has emerged as a promising strategy for similarity learning in MOT applications. Unlike traditional instance similarity-based methods that use only a single positive sample, this new generation of approaches allows training with multiple positive samples, significantly enhancing object similarity learning.

In the next chapter, we will further explore these advancements and discuss how our semi-automatic labeling framework can significantly contribute to improving the video annotation process for MOT, accelerating research in this field.

2.3.2 Multi-Object Tracking

The main challenge in MOT is to accurately track objects over time, even when they traverse complex scenarios with partial occlusion, lighting variations, changes in perspective, and interactions among the objects themselves. Additionally, it is essential to ensure that objects are correctly associated across frames, avoiding errors such as misassigning object identities or losing tracks.

2.3.3 Challenges in Multi-Object Tracking

Multi-Object Tracking presents unique challenges that make it a complex task:

- **Occlusion:** Objects can be obscured by other objects or the surrounding structure, making accurate tracking during occluded periods challenging. Examples shown in 2.2.
- **Lighting Changes:** Lighting variations can significantly alter the appearance of objects, creating challenges in their continuous identification. Examples shown in 2.2.
- **Scale Variations:** Objects can change in size over time, for example, due to their distance from the camera or intrinsic movements. Examples shown in 2.2.

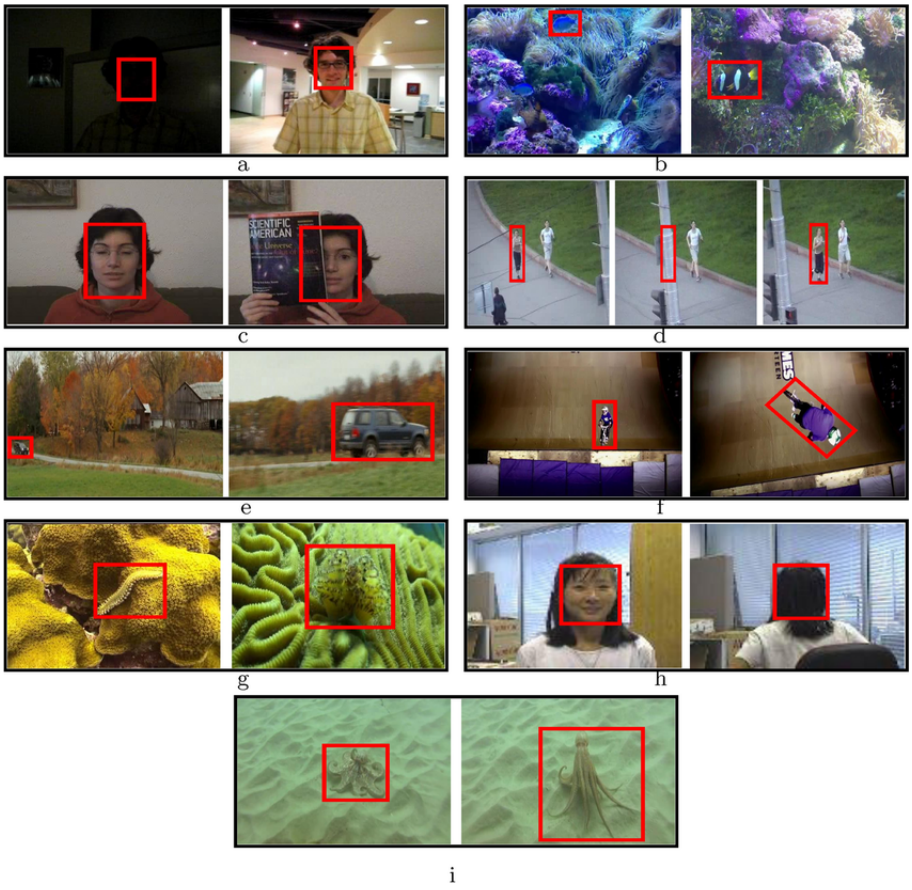


Figure 2.2: Image from [28] Visual illustration of various challenges in object tracking with examples of variations.

Advanced models tackle these challenges through the use of robust tracking techniques, the incorporation of contextual information, and the learning of object motion patterns. However, these difficulties can lead to errors during inference, resulting in the following issues:

False Positives False positives occur when the detection system mistakenly identifies an object in a frame where it is not actually present. These errors can lead to incorrect associations between objects in subsequent frames and compromise tracking accuracy. Reducing false positives is crucial to ensure accurate tracking.

Identity Switches (ID Switches) Identity switches occur when the tracking system mistakenly assigns one object's identity to another object in subsequent frames. This problem is particularly critical in scenarios where objects may be temporarily occluded or separated by other objects. ID switches can lead to serious tracking errors and must be carefully managed.

Recently Appeared Objects In the context of Multiple Object Tracking (MOT), situations may arise where new objects enter the camera's detection area. Managing recently appeared objects is a challenge because the system must determine whether an object previously invisible is the same object detected later. This requires a balance between associating existing objects and recognizing recently appeared objects.

Terminated Tracks Terminated tracks occur when an object previously detected is no longer visible or has exited the camera's detection area. Managing terminated tracks is important to avoid associating objects that are no longer present in the video. Track termination needs to be detected and handled appropriately.

These issues represent some of the key challenges in multi-object tracking and require the implementation of sophisticated algorithms to address them. Success in overcoming these challenges is crucial for achieving accurate and reliable tracking in a variety of application scenarios.

2.3.4 Metrics

In the realm of object detection and tracking, the Clear MOT metrics [4] have emerged as the gold standard for performance evaluation. Conceived with the ambition of presenting a multifaceted examination of a tracking system's efficiency, these metrics span the entire spectrum from pinpointing objects to trailing them across sequences. The Clear MOT metrics chosen for our assessment include:

- **IDP (Identification Precision)**: Captures the consistency in assigning identifiers to objects across consecutive frames.
- **IDR (Identification Recall)**: Measures the system's prowess in recognizing objects without introducing spurious identifiers.
- **RCLL (Recall)**: Highlights the overall effectiveness in detecting objects in specific frames.
- **PRCN (Precision)**: Assesses the precision in object recognition, contrasting true detections against false positives.

- **GT (Ground Truth)**: Quantifies the diversity and count of objects in the dataset to gauge dataset complexity.
- **MT (Mostly Tracked)**: Emphasizes the system's capability to consistently track objects across a significant portion of their presence.
- **PT (Partially Tracked)**: Records instances where objects are intermittently tracked, pointing to potential challenges in tracking.
- **ML (Mostly Lost)**: Notes objects that are seldom tracked, indicating potential difficulties in complex tracking scenarios.
- **FP (False Positives)**: Counts detections of non-existent objects, which can be critical in contexts like autonomous driving.
- **FN (False Negatives)**: Catalogs missed detections, providing insights into potential system limitations.
- **IDs (ID Switches)**: Chronicles instances where tracked objects get newly assigned identifiers, which can compromise tracking consistency.
- **FM (Fragmentations)**: Demonstrates the frequency of objects transitioning between tracked and untracked states.
- **MOTA (Multiple Object Tracking Accuracy)**: An encompassing metric that integrates false positives, misses, and ID switches to deliver a comprehensive view of tracking accuracy.
- **MOTP (Multiple Object Tracking Precision)**: Measures the precision in localizing tracked objects in terms of their spatial accuracy.

These metrics, in tandem, provide a thorough evaluation framework for object detection and tracking systems, aiding in the identification of strengths, weaknesses, and areas for improvement.

2.3.5 State of the Art for MOT

Multi-Object Tracking (MOT) represents one of the fundamental challenges within the field of computer vision. This problem holds great importance in various practical

applications, including security monitoring, autonomous driving, video analysis, and others. In general, MOT aims to identify, track, and monitor multiple objects simultaneously within sequences of images or videos.

Contemporary MOT methods, as mentioned earlier, primarily follow the "tracking by detection" paradigm, which means they detect objects in each frame and subsequently associate them based on the similarity of estimated instances.

Recently, innovative approaches [3, 5, 8, 43] have emerged, demonstrating that accurate object detection, coupled with the spatial proximity measured through Intersection over Unions (IoUs) between objects in consecutive frames, provides a strong indication for object association. This position-based heuristic works well in simple scenarios but can easily lead to errors in the presence of occlusion or crowded scenes. For instance, if objects are occluded or scenes are highly congested, this logic starts to falter. To address such issues, some works [10, 26] have introduced methods for motion estimation to improve accuracy in motion estimation.

This type of approach relies on the use of appearance similarity in the background, making it less effective. The search region used is limited to local proximity to avoid distractions, as appearance features are not effective.

The difference in QDtrack is based precisely on this concept, namely the ineffectiveness of appearance features. In particular, as shown in Figure 2.3, previous methods consider instance similarity learning as a subsequent phase to object detection, thus using only the bounding box of the ground truth, effectively ignoring most of the proposed regions in the image. What is demonstrated is that by "learning" a correct object representation, a search in the embedded space should associate instances without too many issues.

The proposal is that of "quasi-dense similarity learning", which matches hundreds of regions of interest in a pair of images using contrastive learning. The examples produced in this way manage to provide a higher number of both correct bounding boxes and "negative" bounding boxes.

Obviously, the inference process also plays an important role, as it needs to consider all the challenges present in MOT. To address the issue of missing objects, "backdrop" objects are added, and through bidirectional softmax, it ensures that absent objects have low similarity scores and are not erroneously associated.

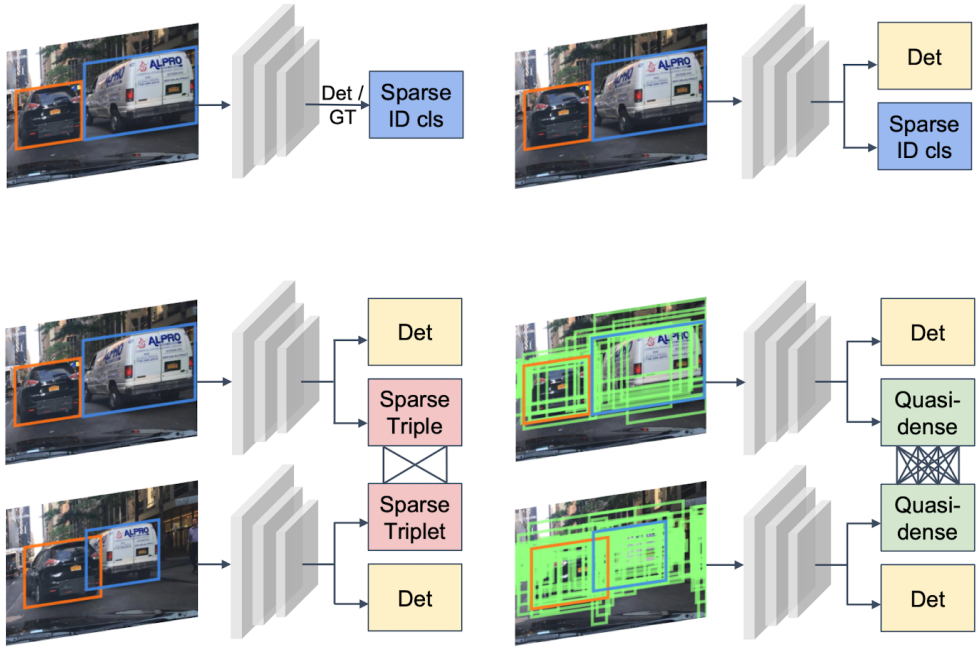


Figure 2.3: Figure from [30]: (a) Traditional ReID model that decouples with detector and learns with sparse ID loss; (b) joint learning ReID model with sparse ID loss; (c) joint learning ReID model with sparse triplet loss; (d) qdtrack quasi-dense similarity learning.

Specifically, the implementation and tests reported in the paper use "Faster R-CNN" [34] as the detector, and despite its simplicity, it has outperformed many methods in the literature. However, we will analyze QDTrack in detail in Chapter 2.5.

2.3.6 The Kalman Filter and its Role in Object Tracking

The Kalman filter is an optimal estimation algorithm for systems with Gaussian uncertainties. Since its introduction in the 1960s [2], it has found application in a wide range of fields, thanks to its ability to operate in real-time and handle noise and uncertainties in measurements.

Fundamental Principles of the Kalman Filter

The Kalman filter is based on a series of predictions and corrections. At each step, the filter makes a prediction about the system's future state and then corrects this prediction when it receives a new measurement.

- **Prediction:** Based on the current state and the system model, the filter makes a prediction about the future state. This prediction will have associated uncertainty, which will increase over time without further measurements.
- **Update (correction):** When a new measurement is received, the filter compares it with the prediction and makes a correction based on the difference between the measurement and the prediction. This update step reduces the uncertainty about the state estimate.

Mathematical Formulation

The Kalman filter is based on a set of mathematical equations that describe the prediction and update of the system state. The key variables are:

1. x: the estimated state of the system.
2. P: the covariance of the state estimate, representing the uncertainty of the estimate.
3. K: the Kalman filter gain, determining how much weight is given to the measurement compared to the prediction.

The prediction equations are:

$$x_{pred} = Ax_{previous} + BuP_{pred} = AP_{previous}A^T + Q \tag{2.3}$$

Where:

1. A is the state transition matrix.
2. B is the control matrix.
3. u is the control vector.
4. Q is the covariance of the process noise.

After the prediction, the update based on the actual measurement is performed:

$$K = P_{pred}H^T(HP_{pred}H^T + R)^{-1}x = x_{pred} + K(z - Hx_{pred})P = (I - KH)P_{pred} \tag{2.4}$$

Where:

1. H is the observation matrix.
2. R is the covariance of the measurement noise.
3. z is the actual measurement.

Application in Object Tracking

In object tracking, the Kalman filter can be used to estimate an object's position and speed in motion. For instance, in a video sequence, if an object moves consistently, the filter can predict where the object will be in the next frame. This is particularly helpful when the object is temporarily occluded or when measurements are noisy.

Advantages and Disadvantages

- **Advantages**:
  - Real-time operation: The Kalman filter is computationally efficient and can operate in real-time.
  - Noise handling: The filter is optimal for systems with Gaussian noise and uncertainty.
- **Disadvantages**:
  - Gaussian assumptions: The filter assumes that the noise and uncertainty are Gaussian, which isn't always true in real-world applications.
  - Linear model: The standard Kalman filter relies on a linear system model. Extensions like the Extended Kalman Filter (EKF) and the Unscented Kalman Filter (UKF) handle non-linear systems.

2.3.7 Multi-Object Tracking (MOT) Techniques

Given the interest in MOT, different approaches have been developed to calculate the similarity between detected objects. In particular, we identify:

Tracking-by-Detection Paradigm In this approach, the process begins with the detection of objects in each frame of the video. Subsequently, the detected objects are associated between consecutive frames based on their instantaneous similarity. This approach is widely used and forms the basis of many contemporary MOT methods.

Utilization of Spatial Proximity and Motion The association of objects in consecutive frames often relies on spatial proximity and the motion of the objects. Spatial proximity can be measured using metrics such as Intersection over Unions (IoU) or the distance between object centers. However, this technique works better in simple scenarios and can lead to errors in complex scenarios with occluded objects or crowded scenes.

Some methods introduce motion estimates, such as the Kalman Filter (a widely used tool described in [2]), optical flow, and motion regression, to improve distance estimates between objects. These methods aim to ensure better association between objects in consecutive frames.

Appearance Similarity The visual appearance of objects, i.e., their appearance, plays a fundamental role in object association and identification. Some approaches focus on the appearance of objects to improve tracking accuracy. These methods use independent recognition models or add feature extraction modules for end-to-end learning.

However, these methods often rely on learning image similarity and measure instantaneous similarity using cosine distance. These approaches can be limited by the number of identities in the training set or the nature of the training triplets used.

2.4 Datasets Used

In this section, we will present the two datasets used for experimentation in the context of the thesis. Both datasets are widely recognized and utilized in the scientific community in the fields of artificial intelligence and computer vision.

2.4.1 BDD100K

To support progress in computer vision applied to autonomous driving, the BDD100K dataset was created, representing a significant step forward in the availability of high-quality visual data. This dataset was designed to overcome the limitations of existing datasets for autonomous driving and was constructed with several key objectives in mind.

Data Volume BDD100K includes an extraordinary collection of over 100,000 videos of diverse driving scenes. These videos capture a wide variety of driving situations in different environments, weather conditions, and roads, thus covering a broad spectrum of possible scenarios that an autonomous vehicle might encounter in the real world.

Geographic and Environmental Diversity A fundamental aspect of BDD100K is its geographic and environmental diversity. This dataset was collected in different locations and lighting conditions, ensuring that models trained on BDD100K are less susceptible to surprises in new conditions.

Complex Annotations To address the challenge of advanced visual understanding required for autonomous driving, BDD100K comes with a comprehensive set of annotations. These include:

- Lane Detection
- Drivable Area Segmentation
- Road Object Detection
- Semantic Segmentation
- Instance Segmentation
- Multi-Object Detection and Tracking
- Domain Adaptation
- Imitation Learning

These detailed annotations enable the study of a wide range of artificial intelligence tasks related to autonomous driving, from lane identification to understanding road objects and tracking multiple objects.

Various Task Complexities BDD100K supports ten different tasks, each with increasing complexity. These tasks range from simple image annotation to the detection and tracking of multiple objects, thus enabling the study of heterogeneous multitask learning. Models can perform a series of tasks with increasing complexity, paving the way for future research on multitask learning algorithms.

Experiments and Benchmarking BDD100K has been used to conduct extensive experiments with the aim of examining the performance of existing algorithms on complex benchmarks. These experiments have highlighted the need for specialized training strategies for existing models to perform heterogeneous tasks.

BDD100K: Our Utilization

BDD100K represents a significant step in providing researchers with a fundamental resource for the development and evaluation of artificial vision algorithms for autonomous driving. This dataset offers a wide range of data and tasks, allowing developers and researchers to tackle complex challenges in this field.

MultiObject Tracking (MOT) We delved into scientific research using BDD100K, focusing specifically on the task of MultiObject Tracking (MOT). To understand the temporal association of objects in videos, we leveraged the multi-object tracking (MOT) dataset, which comprises 2,000 videos with approximately 400,000 frames. Each video has an approximate duration of 40 seconds and is annotated at 5 fps, resulting in about 200 frames per video. Overall, we observed a total of 130,600 track identities and 3.3 million bounding boxes in the training and validation sets. The dataset split is as follows: 1,400 videos for training, 200 videos for validation, and 400 videos for testing.

The tracking benchmark provides data an order of magnitude larger than the previously popular tracking dataset, MOT17 [27]. Even when compared to a similarly recent dataset released by Waymo [37], BDD100K has a greater number of tracking

Dataset	Frames	Sequences	Identities	Boxes
KITTI	8K	21	917	47k
MOT17	34K	21	1638	337k
Waymo	230k	1150	-	9.9M
BDD100K	318K	2k	131K	4.2M

Table 2.1: Table from [42]: Comparison between BDD100K and previous tracking datasets.

sequences (2,000 vs 1,150) and a higher total number of frames (398,000 vs 230,000), in addition to offering greater diversity, including various weather conditions and more geographical locations.

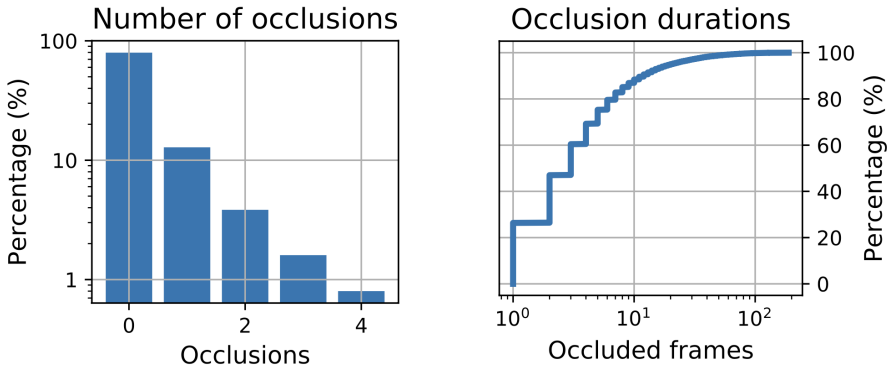


Figure 2.4: Cumulative distributions of the box size (left), the ratio between the max and min box size for each track (middle), and track length (right).

The diversity of the BDD100K MOT dataset is also evident in the object sizes. Furthermore, looking at the cumulative distributions of bounding box sizes, comparing the maximum and minimum sizes of boxes along each track and the lengths of each track, it demonstrates diversity not only in the visual scale between tracks but also in the temporal range of each track.

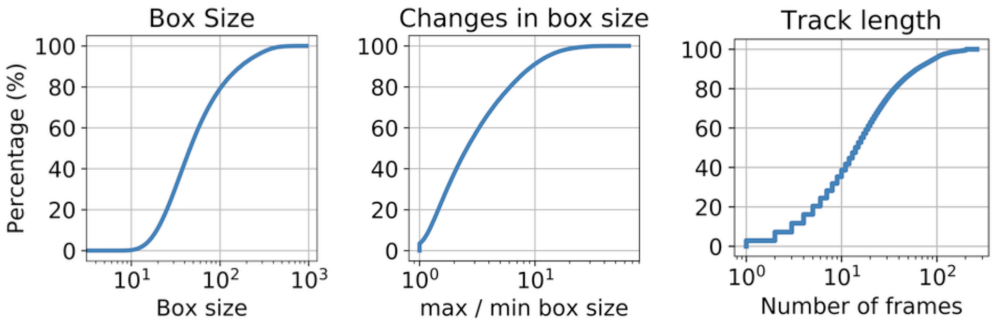


Figure 2.5: Figure from [42]: Number of occlusions by track (left) and the number of occluded frames for each occlusion (right).

Objects in the tracking data exhibit complex patterns of occlusion and reappearance, as shown in Figure 2.5. An object can be completely obscured or exit the frame, only to reappear later. We observed a total of 49,418 occurrences of occlusion in the dataset, corresponding to an occlusion occurrence every 3.51 tracks. The dataset represents the real challenges of object re-identification for autonomous driving tracking.

We leveraged the most advanced scientific research based on this specific dataset, specifically analyzing the MultiObject Tracking task. On this task, the model that achieved the best results, including from the official BDD100K repository, is the QDTrack model analyzed in Chapter 2.5.

2.4.2 Waymo Open Dataset

The Waymo Open Dataset is another significant dataset used in autonomous driving research. It consists of a collection of autonomous driving scenes captured by Waymo vehicles and encompasses a wide variety of driving scenarios, including urban roads, highways, and residential areas. The Waymo dataset is renowned for its vastness and diversity, as it has been collected in various locations and under different lighting and traffic conditions. This makes it a representative dataset of real-world situations that autonomous driving systems may encounter.

This new dataset comprises 1,150 scenes, each extending for 20 seconds, and includes well-synchronized and calibrated high-quality LiDAR and photographic data captured in a range of urban and suburban geographies. The creators have extensively annotated this data with 2D bounding boxes (photographic images) and 3D bounding boxes (LiDAR), with consistent identifiers across frames. Ultimately, it provides a solid foundation for 2D and 3D detection and tracking tasks.

Autonomous driving technology should enable a wide range of applications with the potential to save many human lives, from robotaxi driving to driverless trucks. The availability of large-scale public datasets and benchmarks has significantly accelerated progress in machine perception tasks, including image classification, object detection, object tracking, semantic segmentation, and instance segmentation. To further accelerate the development of autonomous driving technology, the largest and most diverse multimodal autonomous driving dataset to date is introduced,

consisting of images recorded by various high-resolution cameras and sensor readings from multiple high-quality LiDAR scanners mounted on a fleet of autonomous vehicles. The geographical area covered by this dataset is substantially larger than that of any comparable autonomous driving dataset, both in terms of absolute extent and distribution of coverage across geographies. The data has been recorded under various conditions in several cities, specifically San Francisco, Phoenix, and Mountain View, with extensive geographical coverage within each city. The paper demonstrates that the differences in these geographies create a distinct domain gap, offering interesting research opportunities in the field of domain adaptation.

The proposed dataset contains a large number of high-quality manually annotated 3D ground truth bounding boxes for LiDAR data and tightly fitting 2D bounding boxes for photographic images. All ground truth bounding boxes contain tracking identifiers to support object tracking. The dataset contains approximately 12 million LiDAR bounding box annotations and approximately 12 million photographic bounding box annotations, resulting in approximately 113,000 LiDAR object tracks and approximately 250,000 photographic object tracks. All annotations were created and subsequently reviewed by trained annotators using production-level annotation tools.

Currently, the dataset consists of 1,000 scenes for training and validation and 150 scenes for testing, each of which extends for 20 seconds. The selection of test set scenes from a separate geographical area allows us to evaluate how well models trained on our dataset generalize to previously unseen regions.

	KITTI	**NuScenes**	**Argo**	**Waymo**
Scenes	22	1000	113	1150
Ann. Lidar Fr.	15K	40K	22K	230K
Hours	1.5	5.5	1	6.4
3D Boxes	80K	1.4M	993K	12M
2D Boxes	80K	-	-	9.9M
Lidars	1	1	2	5
Cameras	4	6	9	5
Avg Points/Frame	120K	34K	107K	177K
LiDAR Features	1	1	1	2
Maps	No	Yes	Yes	No
Visited Area (km^2)	-	5	1.6	76

Table 2.2: Comparison of datasets.

Ground Truth Labels

The dataset provides high-quality ground truth annotations for both LiDAR sensor data and camera images. The separate annotations for LiDAR and camera data open up new and exciting research opportunities in sensor fusion. For any label, length, width, and height are defined as the dimensions along the x, y, and z axes, respectively.

The dataset comprehensively annotates vehicles, pedestrians, signs, and cyclists in the LiDAR sensor data. Each object is labeled as a 3D vertical bounding box with 7 degrees of freedom (cx, cy, cz, l, w, h, θ) along with a unique tracking ID, where cx, cy, cz represent the center coordinates, l, w, h are respectively the length, width, and height, and α denotes the orientation angle in radians of the bounding box. Figure 2.6 illustrates an example of an annotated scene.

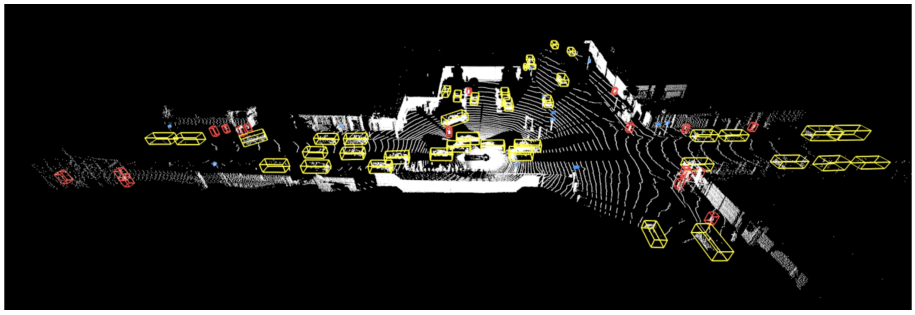


Figure 2.6: Figure from [37]: LiDAR label example. Yellow = vehicle. Red = pedestrian. Blue = sign. Pink = cyclist.

In addition to the LiDAR labels, comprehensive annotations for vehicles, pedestrians, and cyclists have been provided for all camera images. Each object has been labeled with a 2D axis-aligned bounding box with 4 degrees of freedom, which complements the 3D bounding boxes and their 2D amodal projections. The label is encoded as (cx, cy, l, w) with a unique tracking ID, where cx and cy represent the central pixel of the bounding box, l represents the length of the bounding box along the horizontal (x) axis in the image frame, and w represents the width of the bounding box along the vertical (y) axis in the image frame. We use this convention for length and width to be consistent with the 3D bounding boxes. An interesting possibility that can be explored using the dataset is predicting 3D bounding boxes using only camera images.

The paper emphasizes multiple times that all ground truth LiDAR and camera

labels were manually created by highly skilled human annotators using high-level labeling tools. Multiple label verification stages were conducted to ensure a high quality of the labels.

MultiObject Tracking (MOT) The dataset is organized into sequences, each lasting 20 seconds, with data generated by multiple sensors sampled at 10Hz. Furthermore, each object in the dataset is annotated with a unique identifier that is consistent across all sequences.

To evaluate tracking performance, we employ the "clear mot" metric [4]. This metric aims to consolidate various aspects of tracking systems, namely the tracker's ability to detect, localize, and track object identities over time, into a single metric for ease of direct method quality comparison:

$$MOTA = 100 - 100\frac{P_{misses} + P_{falsepositives} + P_{mismatches}}{P_g} \tag{2.5}$$

Here, P_{misses}, $P_{falsepositives}$, and $P_{mismatches}$ represent the counts of missed detections, false positives, and mismatches, respectively, while P_g is the ground truth count.

To calculate $MOTA$, the total number of ground truth objects P_g is used as the denominator. P_{misses} is the sum of counts of all ground truth objects that were not tracked, $P_{falsepositives}$ represents tracked objects without a corresponding ground truth match, and $P_{mismatches}$ indicates mismatches between tracked objects and ground truth objects considered incongruent. The goal is to maximize $MOTA$, which means minimizing the number of missed detections, false positives, and incongruities.

Another metric used to evaluate object tracking is the Multiple Object Tracking Precision (MOTP), measuring the average spatial accuracy of associations between tracked objects and ground truth. It is calculated as:

$$MOTP = 100\frac{\sum_{i,t} d_t^i}{\sum_t C_t} \tag{2.6}$$

Here, P_c represents the total number of associations between tracked objects and ground truth, d_{it}^c is the distance between a tracked detection and the corresponding ground truth, and C_t is the number of associations in time interval t. $MOTP$ quantifies how accurate the tracked objects are compared to the ground truth.

Additionally, other metrics such as Precision, Recall, and F1-score can be employed

to evaluate tracking performance in more detail. Precision measures the percentage of tracked objects that are correct compared to ground truth, while Recall measures the percentage of ground truth objects that have been tracked correctly. The F1-score is a harmonic mean between Precision and Recall, providing an overall measure of tracking quality.

In summary, the use of various metrics enables a comprehensive evaluation of object tracking performance in the dataset, allowing for a detailed assessment of tracking system capabilities.

2.5 QDTrack in Detail

As indicated in the chapter dedicated to the analysis of BDD100K, the QDTrack model has proven to be one of the best in tackling the Multi-Object Tracking (MOT) task on this dataset. In this section, we will delve deeper into the workings of QDTrack and its key characteristics.

The "quasi-dense similarity learning" is proposed to learn an embedding feature space that can associate identical objects and distinguish different ones in online multiple object tracking. "Dense matching" is defined to match candidate boxes for all pixels, while "quasi-dense" indicates that only the most potential candidates are considered. Furthermore, "sparse matching" means that the method only considers "ground truth" labels as matching candidates when learning object associations.

Object Detection The specific implementation uses Faster R-CNN [34] with Feature Pyramid Network (FPN) [22], but it can be easily integrated with other detectors. Faster R-CNN, as previously mentioned in the detector chapter, is a two-stage detector that uses Region Proposal Network (RPN) to generate Regions of Interest (RoIs), then locates and classifies the regions to obtain semantic labels and locations. The entire network is then optimized through a multi-task loss function:

$$\mathcal{L}_{det} = \mathcal{L}_{rpn} + \lambda_1 * \mathcal{L}_{cls} + \lambda_2 * \mathcal{L}_{reg} \tag{2.7}$$

Where $\mathcal{L}_{rpn}$, $\mathcal{L}_{cls}$, and $\mathcal{L}_{reg}$ remain the same as in the original paper.

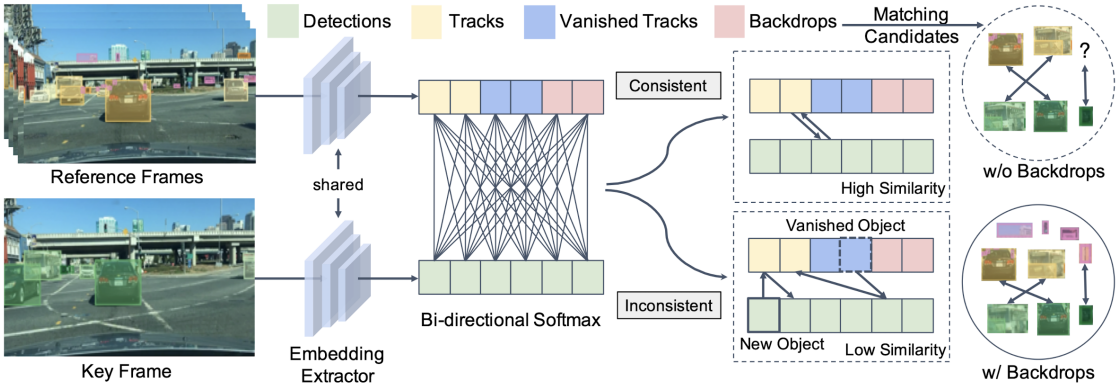


Figure 2.7: Figure from [30]: The testing pipeline of qdtrack method. They maintain the matching candidates and use bi-softmax to measure the instance similarity so that they can associate objects with a simple nearest neighbour search in the feature space

Quasi-dense similarity learning The regions proposed by RPN are used to learn "instance similarity" with quasi-dense matching. What is done is, given a specific image, to select images in a nearby temporal space, where the distance between neighbors is dictated by the parameter K, which in the case of the experiment is $k \in [-3, 3]$.

Once RoIs are generated from the two images and RoI Aligning [19] is used to obtain the feature map, another embedding head parallel to the original bounding box is added, allowing feature extraction for each RoI. The RoIs are defined as positive if the IoU is greater than α_1 or negative if it is less than α_2.

Object Association During the course of experiments, the techniques used in the inference phase have mitigated the tracking-related issues mentioned earlier.

Bi-Directional SoftMax Starting with what is known as "Bi-Directional SoftMax", a technique that bi-directionally compares the embedded space. In Figure 2.7, we can see the testing pipeline. What is done is to use both results obtained by calculating the softmax in both directions in the score function, as shown in Formula 2.8. The higher the score, the closer the two objects are in both vector spaces.

$$f(i,j) = [\frac{\exp(n_i * m_j)}{\sum_{k=0}^{M-1} \exp(n_i * m_k)} + \frac{\exp(n_i * m_j)}{\sum_{k=0}^{N-1} \exp(n_k * m_k)}]/2 \tag{2.8}$$

No Target Cases Objects that have no target, such as new objects and vanished tracks, in the feature spaces should not match with any candidates. Bi-softmax can handle this because it is challenging for them to obtain bidirectional consistency. When an object has high confidence in detection, a new track is started. Furthermore, objects that do not meet this requirement and do not find a match are still utilized and referred to as "backdrops", reducing the number of false positives.

Multi-Target Cases This scenario is handled by many state-of-the-art detectors by removing "intra-class" duplicates through "None Maximum Suppression" (NMS), effectively not addressing cases related to the same object with different classes. To address this issue, "inter-class NMS" is applied, dependent on the intersection over union (IoU) between different areas.

2.5.1 Tested Datasets

In this section, we describe the datasets used to evaluate the performance and reliability of the model. Among the selected datasets, we have the MOT benchmark [27], the new large-scale benchmarks BDD100K [42], Waymo [37], and TAO [13]. These experiments were conducted to assess the performance of QDTrack in various contexts and contribute to future research in the field of multiple object tracking on large-scale datasets.

MOT Challenge: Experiments were conducted on two MOT benchmarks: MOT16 and MOT17 [27]. The dataset comprises 7 videos (5,316 images) for training and 7 videos (5,919 images) for testing. In this benchmark, only pedestrians are evaluated. The frame rates of the videos vary from 14 to 30 FPS.

BDD100K: They used the BDD100K detection training set and tracking training set [42] for training and the tracking validation/test set for testing. BDD100K annotates 8 categories for evaluation. The detection set contains 70,000 images. The tracking set includes 1,400 videos (278,000 images) for training, 200 videos (40,000 images) for validation, and 400 videos (80,000 images) for testing. Images in the tracking set are annotated at a rate of 5 FPS compared to a video frame rate of 30 FPS.

Waymo: The Waymo open dataset [37] comprises images captured from 5

Method	Split	mMOTA ↑	mIDF1 ↑	MOTA ↑	IDF1 ↑	FN ↓	FP ↓	ID Sw. ↓	MT ↑	ML ↓	mAP ↑
Yu et al. [42]	val	25.9	44.5	56.9	66.8	122406	52372	8315	8396	3795	28.1
QDTrack	val	**36.6**	**50.8**	**63.5**	**71.5**	**108614**	**46621**	**6262**	**9481**	**3034**	**32.6**
Yu et al. [42]	test	26.3	44.7	58.3	68.2	213220	100230	14674	16299	6017	27.9
DeepBlueAI	test	31.6	38.7	56.9	56.0	292063	**35401**	25186	10296	12266	-
madamada	test	33.6	43.0	59.8	55.7	209339	76612	42901	16774	**5004**	-
QDTrack	test	**35.5**	**52.3**	**64.3**	**72.3**	**201041**	80054	**10790**	**17353**	5167	**31.8**

Table 2.3: Table from [30]: Results on BDD100K tracking validation and test set. QDTrack method outperforms all methods on this benchmark.

Method	Split	Category	MOTA ↑	IDF1 ↑	FN ↓	FP ↓	ID Sw. ↓	MT ↑	ML ↓	mAP ↑
IoU baseline [23]	val	Vehicle	38.25	-	-	-	-	-	-	45.78
Tracktor++ [3, 23]	val	Vehicle	42.62	-	-	-	-	-	-	42.41
RetinaTrack [23]	val	Vehicle	44.92	-	-	-	-	-	-	45.70
QDTrack	val	Vehicle	55.6	66.2	514548	214998	24309	17595	5559	49.5
QDTrack	val	All	44.0	56.8	674064	264886	30712	21410	7510	40.1
Method	Split	Category	MOTA/L1 ↑	FP/L1 ↓	MisM/L1 ↓	Miss/L1 ↓	MOTA/L2 ↑	FP/L2 ↓	MisM/L2 ↓	Miss/L2 ↓
Tracktor [20, 37]	test	Vehicle	34.80	10.61	14.88	39.71	28.29	8.63	12.10	50.98
CascadeRCNN-SORTv2*	test	All	50.22	7.79	2.71	39.28	44.15	**6.94**	2.44	46.46
HorizonMOT*	test	All	51.01	7.52	2.44	**39.03**	**45.13**	7.13	2.25	**45.49**
QDTrack (ResNet-50)	test	All	49.40	**7.41**	1.46	41.74	43.88	7.10	**1.31**	48.21
QDTrack (ResNet-101 + DCN)	test	All	**51.18**	7.64	**1.45**	39.73	45.09	7.20	**1.31**	46.41

Table 2.4: Table from [30]: Results on Waymo tracking validation set using the py-motmetrics library (top) and test set using official evaluation. * indicates methods using undisclosed detectors.

cameras oriented in 5 different directions: front, front-left, front-right, left, and right. It includes 3,990 videos (790,000 images) for training, 1,010 videos (200,000 images) for validation, and 750 videos (148,000 images) for testing. Three classes are annotated for evaluation. Videos are annotated at a rate of 10 FPS.

TAO: The TAO dataset [13] annotates a total of 482 classes, which are a subset of the LVIS dataset [18]. It comprises 400 videos, with 216 classes in the training set, 988 videos with 302 classes in the validation set, and 1,419 videos with 369 classes in the test set. Classes in the training, validation, and test sets may not overlap. Videos are annotated at a rate of 1 FPS. Objects in TAO follow a long-tail distribution, with half of the objects being people and 1/6 of the objects being vehicles.

2.5.2 Reported Results

The method outperforms all existing methodologies on the above-mentioned benchmarks without the use of any specific tricks. Performance is evaluated using official metrics.

MOT Results achieved with private detectors on the MOT16 and MOT17 benchmarks are shown in Table 2.5. The model achieves the highest MOTA of 68.7% and an IDF1 of 66.3% on MOT17. It surpasses the state-of-the-art tracker, CenterTrack

Dataset	Method	MOTA ↑	IDF1 ↑	MOTP ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDs ↓
MOT16	TAP [44]	64.8	**73.5**	78.7	292	164	12980	50635	**571**
	CNNMTT [24]	65.2	62.2	78.4	246	162	6578	55896	946
	POI*[41]	66.1	65.1	**79.5**	258	158	**5061**	55914	3093
	TubeTK_*POI** [29]	66.9	62.2	78.5	296	**122**	11544	47502	1236
	CTrackerV1 [32]	67.6	57.2	78.4	250	175	8934	48305	1897
	QDTrack	**69.8**	67.1	79.0	**316**	150	9861	**44050**	1097
MOT17	Tracktor++v2 [3]	56.3	55.1	78.8	498	831	**8866**	235449	1987
	Lif_T*	60.5	65.6	78.3	637	791	14966	206619	**1189**
	TubeTK*[29]	63.0	58.6	78.3	735	**468**	27060	177483	4137
	CTrackerV1 [32]	66.6	57.4	78.2	759	570	22284	160491	5529
	CenterTrack*[43]	67.8	64.7	78.4	816	579	18498	160332	3039
	QDTrack	**68.7**	**66.3**	**79.0**	**957**	516	26589	**146643**	3378

Table 2.5: Table from [30]: Results on MOT16 and MOT17 test set with private detectors. Note that we do not use extra data for training. ↑ means higher is better, ↓ means lower is better. * means external data besides COCO and ImageNet is used.

[43], by 0.9 points in MOTA and 1.6 points in IDF1, respectively. The method does not achieve a relatively low ID Sw. because we have a higher recall.

BDD100K Key results on the validation and test sets of the BDD100K tracking are reported in Table 2.3. The mMOTA and mIDF1, representing object coverage and identity consistency, are 36.6% and 50.8% on the validation set and 35.5% and 52.3% on the test set. In both sets, the method surpasses the baseline benchmark reference by 10.7 points and 9.2 points in terms of mMOTA and 6.3 points and 7.6 points in terms of mIDF1, respectively. It also outperforms the BDD100K 2020 MOT Challenge baseline (madamada) by a significant margin but with a simpler detector. The significant improvements demonstrate that the method enables more stable object tracking.

Waymo Table 2.4 shows the main results on the Waymo open dataset. We report results on the validation set following the configuration of RetinaTrack [23], which experiments only on the vehicle class. The method outperforms all baselines on both the validation and test sets. They achieve a MOTA of 44.0% and an IDF1 of 56.8% on the validation set. They also obtain a MOTA/L1 of 49.40% and a MOTA/L2 of 43.88% on the test set. Vehicle performance on the validation set is superior by 10.7, 13.0, and 17.4 points compared to RetinaTrack [23], Tracktor++ [3, 23], and IoU baseline [23], respectively. The model with ResNet-101 and deformable convolution (DCN) achieves state-of-the-art performance on the test benchmark, which is on

par with the Waymo 2020 2D Tracking Challenge winner (HorizonMOT) but with a simple single-model setup.

TAO The model achieves 16.1 points and 12.4 points of AP50 on the validation and test sets, respectively. The results are superior by 2.9 points and 2.2 points compared to the solid TAO baseline, which is 13.2 points and 10.2 points, respectively. Although it only slightly improves the overall performance by 2-3 points, we observe that it significantly outperforms the baseline on frequent classes, namely 38.6 points against 18.5 points on pedestrians.

2.6 Video Annotation Tools

The creation of labeled datasets is a fundamental element for training artificial intelligence models. This section will explore the various solutions and tools available for video annotation, focusing in particular on platforms that support semi-automatic labeling, a crucial aspect for optimizing the efficiency of this process, as well as understanding the State of the Art.

In the context of video annotation, it is essential to address the challenges related to recognizing and labeling objects and actions within video sequences. This task is significantly more complex than annotating static images, as it involves tracking moving objects across thousands of frames, each of which may contain relevant information. The following tools have been selected for their ability to simplify the video annotation process, especially through the use of semi-automatic techniques.

2.6.1 Video Annotation Tools

iVAT

iVAT [6] is a powerful interactive video annotation tool based on C/C++ libraries and uses Qt libraries for the user interface, along with the Open Computer Vision Library (OpenCV) for computer vision algorithms. This open-source tool offers considerable flexibility due to its platform independence, allowing it to be used on mobile devices as well.

The annotation analysis feature of iVAT manages each video annotation session

as a separate project. Additionally, iVAT includes an analysis module that segments videos into individual shots using scene detection algorithms. This further simplifies the annotation process by enabling users to focus on specific sequences within the video.

ViTBAT

ViTBAT [7] is a reliable video annotation tool that focuses on generating ground-truth information for individual tracking and group-level behavioral analysis. Implemented in MATLAB, it utilizes the MATLAB Computer Vision toolbox.

ViTBAT adopts a state-based and behavioral approach at both the individual and group levels within videos. This methodology proves particularly effective for annotating actions and interactions within video sequences. The use of point-based or area-based representations offers significant flexibility in labeling objects and behaviors within the video.

MViPER-GT

MViPER-GT [35] represents an evolution of ViPER, an open-source project based on Java supported by the ViPER API. This tool distinguishes itself with its ability to automatically track objects within videos and calculate non-linear trajectories for each object or track in the video frame.

Once a session's annotation is completed, MViPER-GT offers a ground-truth control that includes a tracking system to enhance real-time video analysis. This video annotation module is particularly suitable for projects requiring accurate traceability and the generation of high-quality training data.

BeaverDam

BeaverDam [36] is a video annotation tool designed for the detection and tracking of targets in crowded and dynamic scenes. This tool primarily focuses on annotating individual targets and does not support group tracking. BeaverDam builds upon concepts previously developed by VATIC and is designed to significantly simplify the setup and installation process compared to its predecessor.BeaverDam offers a web-based interface that greatly surpasses VATIC's command-line-based interface

[39], making BeaverDam an efficient and user-friendly cloud-based video annotation tool.

VATIC

VATIC [39] is a well-known large-scale video annotation tool that relies on high-quality video annotations. This tool leverages crowdsourcing but distinguishes itself with the ability to identify a small but expert group of workers capable of delivering high-quality results.

A distinctive feature of VATIC is the ability for users to add multiple attributes to annotations, in addition to the ability to advance and rewind in videos, adjust playback speed, and focus on objects of interest. These additional features make VATIC a powerful option for projects requiring high precision.

CVAT - Computer Vision Annotation Tool

CVAT [11] is an interactive tool for video and image annotation used in the field of computer vision. It is used by tens of thousands of users and companies worldwide. The mission is to assist developers, businesses, and organizations around the world in solving real-world problems using a data-centric AI approach.

CVAT has an online version that can be used for free or by subscribing to get unlimited data, organizations, auto-annotations, and integrations with Roboflow and HuggingFace. Additionally, you can set up CVAT as a self-hosted solution following the self-hosted installation guide. It offers Enterprise support for self-hosted installations with premium features, including SSO, LDAP, integrations with Roboflow and HuggingFace.

2.7 Development of a New Semi-Automatic Labeling Framework

A fundamental step in the scope of this thesis involves the development of an innovative framework for semi-automatic labeling, specifically designed to leverage specialized models in autonomous driving, such as QDTrack. This new framework

primarily focuses on Multi-Object Tracking (MOT) activities and is customizable to meet the specific requirements of any other network used in this context.

The primary goal of this framework is to combine the most advanced deep learning techniques for object tracking and recognition to optimize the annotation process. This will allow human operators to interact with the system more efficiently and intuitively, minimizing the manual intervention required to correctly label objects within videos.

One of the distinctive features of this framework is its flexibility. It has been designed to be highly customizable, enabling users to tailor it to the specific needs of their autonomous driving projects. This means it can be successfully used with a wide range of neural networks and datasets, making it an extremely versatile tool for video annotation.

The framework, although developed to work with QDTrack, will be capable of integrating existing tracking models, with the ability to make a few modifications to adapt to any other network used in this context, making it highly versatile.

In summary, the development of this new semi-automatic labeling framework represents a significant step towards optimizing and accelerating the video annotation process for autonomous driving. Its customization and adaptability to a variety of networks make it a powerful tool for improving the efficiency and quality of video labeling, thus contributing to the advancement of research in this field.

Chapter 3

Original contribution to problem solution

3.1 Original Contribution to Problem Solution

In this chapter, we will present our original contribution to the problem of object labeling in video for autonomous driving. We will describe the proposed methodology, methodological innovations, system design, technological and application innovations, as well as the tools, technologies, and models used to realize our contribution.

As described in the previous chapters, annotating a dataset is a fundamental and highly time-consuming process. The innovation in this thesis work aims to drastically reduce annotation times by leveraging the potential of transfer learning and deep learning. Transfer learning is a technique that allows transferring knowledge learned from a pre-trained model on a source domain to a new target domain, thereby reducing the need for labeled data. Deep learning is a subfield of artificial intelligence that relies on multilayer artificial neural networks capable of automatically learning hierarchical data representations.

The Goal Our goal is to utilize a pre-trained model called QDTrack, which can perform object detection, classification, and tracking in videos, to annotate another dataset or, more precisely, to enrich the accompanying information.

The Idea The idea is to apply QDTrack to the Waymo dataset, using ground truth as supervision to adapt the model to the new domain. This way, we can obtain automatic predictions for labeling objects in videos without having to manually annotate each frame. However, QDTrack's predictions may not be perfect and may contain errors or inaccuracies. For this reason, we have developed a set of algorithms that, whether or not they use ground truth, can improve QDTrack's predictions and reduce the associated errors.

The Algorithm The goal of the developed algorithms is to further optimize the predictions by addressing issues such as ID Switch, ID Merge, and other related challenges. ID Switch refers to the case where two objects mistakenly exchange their identifiers during the video, causing a tracking discontinuity. ID Merge refers to the case where two distinct objects are merged into one by the model, resulting in a loss of information about their identity and position. These problems can arise due to various factors such as occlusion, overlap, or object similarity. After resolving the aforementioned issues and replacing different track IDs, it is necessary to ensure that the associated class for that track ID remains consistent throughout the video. To achieve this, another algorithm has been developed, closing the loop.

3.2 Definition of the Proposed Methodology

In this section, we will provide a clear definition of the methodology we have developed to address the problem of object labeling in videos for autonomous driving. Our methodology has been designed to effectively and efficiently tackle the challenges related to dataset annotation. We will describe the approaches, strategies, and key methods we have employed to successfully meet this challenge.

3.2.1 General Approach

Our methodology's approach is structured into several interconnected phases, with the aim of achieving accurate and efficient object labeling in videos. These phases include:

1. **Data Acquisition**: We draw data from the Waymo dataset, rich in autonomous

driving data, selecting images from the front camera and extracting information related to bounding boxes and tracking IDs (trackID).

2. **Data Pre-processing**: Ensuring consistency between ground truth data and those used during the inference phase.

3. **Inference**: Using prepared images for object detection and tracking.

4. **Label Enhancement**: Applying advanced algorithms to further optimize labels and address common tracking issues, including:

   (a) ID Merge:

   Resolving ID Merge problems where two distinct objects may erroneously share the same tracking ID. This step is crucial for ensuring accurate object identification.

   (b) ID Switch:

   Addressing cases where tracking IDs have been mistakenly exchanged between different objects during the tracking process. Resolving this situation contributes to maintaining correct correspondence between objects and temporal tracking.

   (c) Class Correction:

   After consolidating tracking IDs, we tackle the task of updating the classes associated with each object. This phase is necessary because previous algorithms primarily focus on tracking IDs and may not account for object classes. Maintaining the correct classes is essential for proper data interpretation.

   (d) Handling "Temporal Gaps":

   In the final phase of the process, we focus on managing "temporal gaps", which are moments when the network is unable to detect an object. These "gaps" can occur in two main situations:

      i. Complete Occlusion:

      If an object is completely obscured by another object (e.g., trees or vehicles with different tracking IDs), we maintain the "temporal gap" in the track since the object is not visible to the detector.

ii. Partial Occlusion:

When an object is only partially obscured, we attempt to "fill" the temporal gap, meaning identifying the object even during periods of partial obscuration. This step is crucial for ensuring continuous and accurate tracking. Additionally, we also try to "extend" the obtained track by searching for the object in the frames before and after its actual detection.

3.3 Methodological Innovations

In this section, we will highlight the main methodological innovations introduced in our approach to object labeling in video for autonomous driving. We will describe in detail how these innovations have contributed to improving efficiency, accuracy, and other crucial aspects of the process.

Reduction in Labeling Times One of the primary objectives of our methodological innovations has been to drastically reduce the time required for data labeling. Our strategy is based on a combination of intelligent automation and targeted human intervention. The idea of a semi-automatic labeling system is precisely to allow annotators to verify the work done, trying to minimize manual adjustments to the labels produced. We have developed advanced algorithms that make the most of available information to automatically label objects when possible, leaving people to analyze complex or ambiguous cases. This approach has led to a significant overall process efficiency improvement, minimizing repetitive manual work.

Automatic Error Correction We have introduced an automatic error correction system that identifies and autonomously resolves many common errors in object tracking. It automatically detects and corrects errors such as ID merge and ID switch, enhancing label consistency and accuracy.

The methodological innovations described above represent the fundamental pillars of our approach to object labeling in video for autonomous driving. Through these methodologies, we have achieved a significant improvement in performance and overall process efficiency.

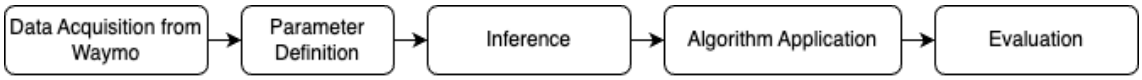


Figure 3.1: Flowchart of the object labeling process in videos using the Waymo dataset. The diagram illustrates the key stages, from initial data acquisition and parameter setup to the inference phase, algorithm application, and results evaluation.

3.4 Design of the Proposed System

In this chapter, we will delve into the overall design of the system developed to address the problem of object labeling in video for autonomous driving. We will provide a detailed analysis of the general architecture, key components, and workflow of the system. We will also describe how data is managed, how algorithms are integrated into the process, and how the system is structured globally.

3.4.1 General System Architecture

The system's architecture has been designed to ensure maximum flexibility and scalability, allowing it to effectively handle various types of data and situations. The system is divided into several interconnected modules that collaborate to perform the various stages of the labeling process.

In essence, to make the proposed algorithms work, it is necessary to model the output data of any MOT (Multiple Object Tracking) model to be consistent with those used in this work. To use this approach for labeling the entire Waymo dataset, you need to follow the flowchart shown in Figure 3.1. In particular:

1. **Data Acquisition from Waymo:** We have developed a dedicated framework for extracting data from Waymo and formatting it for use with our system. The data provided by Waymo is stored in .parquet files, which represent a highly efficient tabular storage format for structured data. These .parquet files contain detailed information about driving scenes captured by Waymo vehicles, including data such as GPS coordinates, LiDAR information, images captured by cameras, and ground truth annotations. The data is organized into various folders within the Waymo dataset. One of these folders contains images captured by vehicle cameras, while others contain ground truth organized for various labeling tasks, such as multi-object tracking (MOT). The ground truth

includes detailed information about objects in the scenes, such as vehicles, pedestrians, traffic signs, and cyclists, along with information about their movements and spatial relationships. Our framework is designed to efficiently extract and manage this data, ensuring proper formatting for subsequent stages of the process, such as inference and the application of label improvement algorithms. This way, we can fully leverage the information provided by Waymo to address the object labeling challenge in autonomous driving videos.

2. **Parameter Definition:** Before running the framework, it is necessary to configure the parameters that influence the process. These parameters include the path to the model checkpoint used, the output path for generated results, the path to images and ground truth, as well as coefficients and other customizable parameters to adapt the algorithms to specific requirements. These parameters can also be differentiated based on the amount of ground truth to be used. In fact, as described in Chapter 3.5.1, we have developed different algorithms based on the quantity of available ground truth.

3. **Inference:** During the inference phase, using the configured parameters, including the one specifying which ground truth information to use, predictions are made. This phase can vary significantly depending on the chosen mode, as described in Chapter 3.5.1. During this phase, data structures are created that will be used in subsequent algorithms.

4. **Algorithm Application:** In this phase, we apply the algorithms we have developed. As mentioned earlier, the choice of algorithms varies depending on the amount of ground truth used, as described in Chapter 3.5.1.

5. **Results and Evaluation**: Once the application of the chosen algorithms is completed, it is essential to conduct a thorough evaluation of the results obtained. For this purpose, we use recognized metrics like the cler MOT to assess the effectiveness of tracking and the classification metric to analyze the accuracy in identifying categories. The choice of these metrics ensures a deep understanding of the performance of the selected method and provides a solid foundation for further optimizations or revisions.

3.5 Methodology and Approaches

After introducing the goal of using QDTrack and the idea of applying it to the Waymo dataset, this section will focus on the specific approaches used for labeling and performance evaluation.

3.5.1 Different implementation

The Waymo dataset provides detailed ground truth, which has been leveraged in various ways depending on the chosen approach.

CompleteGroundTruth

In the *CompleteGroundTruth* approach, labeling fully benefits from the detailed information provided by the ground truth. Although the predicted class is the only element taken from the model, when available, the process unfolds in several crucial stages:

- **Inference Phase**: After obtaining predictions for a specific image, a comparison is made between all bounding boxes, associated with a trackID and a class, and those present in the ground truth. Intersection over Union (IoU) serves as the metric for this comparison. In the presence of a significant match, the ground truth bounding box is adopted along with the trackID, while keeping the predicted class. Ground truth bounding boxes without matches are instead labeled with a generic class, which will be further refined by subsequent algorithms with the goal of removing it and assigning the object an actually predicted class by QDTrack.

- **Class Fixing**: This phase focuses on correcting predicted classes. Since the trackID and BBOX come from the ground truth, the main challenge lies in inconsistencies between predicted classes in different frames for the same object. The goal is to ensure that a given object is correctly classified across different frames.

  To do this, we analyze the predicted classes, their confidence, and the sizes of bounding boxes for a given trackID. The formula we apply is as follows:

$$normFactor = width \times height \tag{3.1}$$

$$score = coeffConfidence \times confidence + coeffBbox \times \frac{bboxarea}{normFactor} \tag{3.2}$$

Where *confidence* represents the confidence with which a specific class was predicted.

The idea behind this formula is to give weight to both the prediction's confidence and the bounding box's area. If a bounding box is large, it indicates that the network was able to clearly identify the object, making the prediction more reliable. As a result, a larger bbox is weighted more in the formula, but without neglecting the prediction's confidence.

What we do is analyze each *trackID* to determine the associated class and confidence. We use the coefficients *coeffConfidence* and *coeffBbox* to assign more or less importance to the confidence or the bbox area. Then, an overall score is calculated for each predicted class for that *trackID*. In the end, the class with the highest overall score is assigned to the respective *trackID*. This process ensures that each object is correctly classified across all frames in which it appears.

Only BBOX

The *OnlyBBOX* approach primarily focuses on utilizing bbox information. This entails several phases and challenges:

- **Inference Phase**: Similar to the previous approach, we compare the predicted bounding boxes with those from the groundTruth using IoU. Bounding boxes that have a match are retained with the predicted class and trackID. Conversely, those without matches are labeled generically and will be handled in subsequent algorithms.
- **ID Switch Fix**: This phase is crucial to ensure consistency among objects detected in consecutive frames. To handle ID Switch issues, where a specific

object may be associated with more than one trackID during the video, the following strategies are adopted:

1. We use predictions based on the Kalman filter to track the expected movement of an object across frames.
2. If a trackID is not present in a frame, we predict its position and compare it with existing bounding boxes in the frame.
3. If we find a high IoU value with a bounding box associated with a different trackID, we identify a potential ID Switch. Otherwise, if no match is detected for a certain number of consecutive frames, we terminate the search for that object.
4. Additionally, to correct potential errors that may have occurred in the early stages of the video sequence, we adopt a retrospective approach. Starting from the last frame where the object was identified, we backtrack to the earlier frames using the same Kalman filter and IoU-based logic.

- **Class Fix**: This phase aims to resolve inconsistencies in the predicted classes by associating each object with the most appropriate class, using the metrics described earlier.

NoGroundTruth

The *NoGroundTruth* approach stands out significantly from others as it relies entirely on model predictions, operating without any information provided by ground truth. This complete independence from ground truth poses unique challenges in labeling, necessitating advanced methodologies to ensure the accuracy and consistency of the generated labels.

- **Fix for ID Merging**: One of the most common issues in this approach is ID merging, where two or more distinct objects are erroneously identified as one, receiving a single trackID. This phenomenon can occur in various circumstances, such as when objects are very similar or the detected bounding boxes are so small that the model cannot extract sufficiently discriminative features. Analyzing these cases, it has been observed that this behavior often

occurs for objects on the video's periphery and for very small bounding boxes, which do not allow for the extraction of distinctive features. Therefore, similar but often distant objects are assigned the same trackID. For example, a trackID located on the left side of the video may be reassigned to the right side after just one frame. These behaviors have been identified as "jumps" and are managed by separating the two entities into different trackIDs. In Figure 3.2, the algorithm in action is shown.

- **Fix for ID Switching**: Similar to the *OnlyBBOX* approach, reliance is placed on the Kalman filter to monitor the trajectory of objects across consecutive frames. This helps identify cases of ID switching, where an object may receive different trackIDs during its appearance in a video. When an ID switch is detected, corrections are applied to ensure that the object maintains a consistent trackID. In Figure 3.3, the algorithm in action is shown.

- **Fix for Classes**: As with other uses of ground truth, problems related to classes are addressed. This phase aims to stabilize class predictions. By analyzing the predicted classes, their confidence, and the dimensions of bounding boxes for each trackID, the most probable class is determined and assigned consistently. In Figure 3.3, the algorithm in action is shown.

- **Fix for Temporal Gaps**: This phase represents one of the major challenges of the *NoGroundTruth* approach. In the absence of ground truth information, tracking objects that disappear and reappear in the video becomes particularly complex. To address this problem, we have developed an algorithm that combines model predictions with the Kalman filter. This algorithm seeks to "fill" temporal gaps, extending object tracking even in frames where they are not directly visible, ensuring continuous tracking. In Figure 3.4, the algorithm in action is shown.

This approach represents the pinnacle of the semi-automatic labeling challenge since it relies exclusively on the model's effectiveness and post-processing techniques to produce accurate results. The results presented in the above images are just an example of what has been achieved and represent scenarios that have become

apparent during the analysis. They are, therefore, a simplified representation of what has been accomplished and the improvements made by our algorithms.

3.5.2 Challenges and Methodological Considerations

The main challenges faced are related to MOT, as discussed in the previous chapters. These challenges include issues such as ID merging, ID switching, and temporal gaps. In the following pages, we provide an example for each scenario to further clarify the concept and how our algorithm has addressed these criticalities:

Figure 3.2: Fix IDMerge Scenario

Figure 3.3: Fix IDSwitch and Fix Classes Scenario

Figure 3.4: Temporal Holes Fix Scenario

Now, let's go into detail and explain further:

Case IDMerge The present case, called IDMerge, is illustrated in Figure 3.2. As clearly highlighted in the upper visual representations, trackID 9 undergoes a significant transformation, moving to the right side of the video. This phenomenon is due to the excessively small size of the bounding box; the extracted features are not distinctive, causing the model to make errors by assigning this trackID to other elements. However, after the implementation of the algorithms, as visible in the lower images, the problem is effectively resolved. It is also noteworthy that this management methodology has contributed to solving an ID switch issue related to

the same trackID 9, which, in the last frame, takes on the identifier 13.

Case IDSwitch The case known as IDSwitch, described in Figure 3.3, is detailed as follows: In the upper images, we observe the object identified as trackID 11, initially classified as a "bus." Subsequently, during the video, the identifier undergoes a transition, changing to "15", accompanied by a class change to "car." This example represents a typical scenario in which a change in the object's class results in the assignment of a new trackID, generating a significant number of ID switches. However, in the lower images, after the application of the algorithms, this behavior no longer occurs.

Case Fix Classes The case known as Fix Classes, similar to the previous IDSwitch case, is illustrated in Figure 3.3. In this case, it is noteworthy that the class associated with a particular object remains constant throughout the entire video duration. This is made possible through the application of our algorithms, which calculate and replace the class associated with each trackID in all results, ensuring stability in classification throughout the tracking process.

Case Temporal Gaps The case under consideration, called "Temporal Gaps", is illustrated in Figure 3.4. From this representation, an interesting behavior regarding trackID 19 emerges. While in the upper images, the object with trackID 19 is detected starting from frame 42, thanks to the implementation of our temporal gap correction algorithm, we were able to anticipate its detection, identifying it as early as frame 5. This highlights the algorithm's ability to fill temporal gaps and ensure early and consistent object tracking along the video sequence.

3.6 Technological and Application Innovations

In the context of computer vision, data labeling plays a crucial role in training and validating machine learning models. However, the annotation process can be extremely time and resource-intensive, especially when dealing with large datasets.

The primary goal of this work was to introduce a semi-automatic labeling system aimed at simplifying and speeding up the annotation process. This innovation has

the potential to revolutionize how annotators interact with datasets, significantly reducing manual workload and associated errors.

In particular, the proposed system leverages the capabilities of MOT (Multiple Object Tracking) models to generate preliminary annotations, which can then be easily reviewed and corrected by annotators. This approach combines the best of both automatic and manual methods, ensuring a high level of precision while maintaining a significant degree of automation.

The use of a pre-trained model on BDD100K further strengthened the robustness and reliability of the proposed system, as BDD100K is one of the reference datasets in the field of object recognition and tracking in images for autonomous driving.

Furthermore, the choice to use the Waymo dataset, one of the largest and most complex datasets available for autonomous driving, underscores the importance and relevance of this innovation. Manually annotating a dataset of such magnitude would be a monumental task, and the introduction of a semi-automatic system represents a significant step forward in managing large datasets.

The innovation proposed in this work not only makes the annotation process more efficient but also opens the door to new possibilities in the field of computer vision, where the quantity and quality of annotated data are crucial for the success of machine learning models.

3.7 Technologies and Models Used for Implementation

We will provide an overview of the specific tools, technologies, and models we used to implement our solution. This will include software libraries, hardware, machine learning frameworks, or any other elements relevant to the realization of our contribution.

3.7.1 Libraries and Frameworks

The core of our solution relies on a set of well-established libraries and frameworks in the field of artificial intelligence and machine learning. Here is a description of the most relevant ones:

- **torch**: A deep learning framework that allowed us to use neural network

models.

- **torchvision**: Provided us with tools for working with images and videos, including pre-trained models, datasets, and transformation methods.
- **mmcv-full** and **mmdet**: These libraries were crucial for our computer vision projects, particularly for object detection and segmentation.
- **onnxruntime**: Enabled us to run models in ONNX format, ensuring interoperability and performance optimization.
- **opencv-python**: A fundamental library for computer vision, providing us with image processing and computer vision capabilities.
- **scikit-learn**: Used for various machine learning algorithms and analysis tools.
- **scipy**, **pandas**, and **numpy**: These libraries played a central role in data manipulation, processing, and analysis operations.
- **seaborn** and **matplotlib**: Assisted us in data visualization and graph generation.

3.7.2 Hardware and Other Tools

During the development of our project, we relied on a combination of hardware resources, including the advanced servers of the university and my local workstation. This combination enabled effective parallel processing, optimizing times and ensuring maximum efficiency in both development and testing phases. From a software perspective, I primarily used VSCode, which provided a versatile and feature-rich development environment for writing and testing code.

3.8 Chapter Conclusion

In this chapter, we have explored in detail the methodology and approaches adopted for semi-automatic video labeling, emphasizing the importance of combining advanced deep learning techniques and the intelligent use of ground truth information provided by the Waymo dataset. We have also discussed the key tools and technologies used in the process, including the hardware and software environments that supported

the development and realization of our project. With a clear understanding of the methodology and tools employed, we can now proceed to the subsequent phases of evaluation and experimentation, where we will assess the effectiveness and efficiency of our solution in the context of video labeling for autonomous driving.

Chapter 4

Experimental validation

4.1 Experimental Validation and Practical Aspects

In the context of research and development, experimentally validating algorithms and understanding their practical aspects is crucial. This not only provides a solid foundation for assessing the effectiveness of proposed solutions but also offers a clear indication of how these solutions would perform in real-world scenarios.

4.1.1 Definition of the Experimental or Verification Protocol

Performance evaluation is not a mere theoretical exercise but rather a milestone in the evolution of any algorithm or system. It is through this detailed analysis that we can truly probe the depth of our solutions and comprehend their potentials. Our choice of metrics was not arbitrary; we relied on established and recognized methods in the scientific community, ensuring that our evaluations are both relevant and comparable to other work in the field.

Clear MOT Metrics

Clear MOT metrics [4] represent a de facto standard in evaluating tracking solutions. They have been designed with the intention of providing a detailed and comprehensive overview of a tracking system's performance, allowing analysis of every aspect, from object detection to their tracking over time. The Clear MOT metrics we have adopted for our analysis are:

- **IDP (Identification Precision)**: Identification precision assesses how often a system correctly assigns the same identifier to an object across consecutive frames. A high value in this metric indicates that the system rarely makes errors in assigning new identifiers to already known objects.
- **IDR (Identification Recall)**: This metric, representing identification recall, measures how often a system correctly recognizes and tracks objects without introducing unnecessary or incorrect new identifiers.
- **RCLL (Recall)**: This metric provides an overview of the system's effectiveness in recognizing objects present in a specific frame. If a system has a high RCLL value, it indicates that it is highly likely to correctly identify the present objects.
- **PRCN (Precision)**: Measures the accuracy with which the system recognizes objects by comparing correctly detected objects with falsely identified objects (false positives). A high value suggests that the system is very precise in distinguishing real objects from possible artifacts or noise.
- **GT (Ground Truth)**: Offers a quantitative perspective on the variety of objects present in the dataset. The higher the count, the more diverse and numerous the objects the system has analyzed, indicating greater dataset complexity.
- **MT (Mostly Tracked)**: This metric highlights the system's ability to track objects for most of their duration. If a system has a high MT value, it means it can maintain a consistent track of objects even in dynamic or complex situations.
- **PT (Partially Tracked)**: Indicates the number of objects that have been tracked only for a portion of their overall duration. High values may suggest challenges in maintaining consistent tracks in scenarios with frequent occlusions or rapid changes.
- **ML (Mostly Lost)**: Evaluates how many objects have been tracked only briefly. High values may indicate issues in tracking objects in complex scenarios or in the presence of strong disturbances.

- **FP (False Positives)**: This metric counts how many times the system detected an object that did not actually exist in the frame. Each false positive can have negative impacts on autonomous driving system performance, making it essential to minimize this value.

- **FN (False Negatives)**: Measures the instances where the system missed an actually present object. These errors can be due to occlusions, noise, or algorithmic limitations.

- **IDs (ID Switches)**: This count represents the times when the system changed the identifier of a tracked object. Situations like intersections or overlaps can lead to undesired ID changes, affecting tracking consistency.

- **FM (Fragmentations)**: This metric indicates how many times an object transitioned from a "tracked" state to an "untracked" state. High values may suggest issues in tracking continuity or occlusion handling.

- **MOTA (Multiple Object Tracking Accuracy)**: A key metric that combines false positives, missed detections, and ID changes to provide an overall view of tracking effectiveness. It considers all aspects of tracking, giving a global indication of system accuracy.

- **MOTP (Multiple Object Tracking Precision)**: Provides an indication of how accurately the system localizes tracked objects. It evaluates how close the predicted tracking is to the actual object position.

These metrics provide a comprehensive and detailed overview of our system's behavior, allowing us to identify both strengths and areas for potential improvement.

4.1.2 Classification Metrics

In addition to the Clear MOT metrics, we deemed it essential to complement our analysis with standard classification metrics. This is because tracking is not just about identifying objects over time but also the ability to classify them correctly. Classification metrics help us evaluate the accuracy and reliability of our predictions. The metrics we have chosen for our analysis are:

- **Accuracy**: This metric provides an overview of the overall effectiveness of a model by measuring the proportion of correctly made predictions compared to the total predictions. It is calculated as the ratio of the total number of correct predictions to the total number of predictions made. While accuracy can offer a general indication of a model's performance, it may not be suitable in scenarios with imbalanced class distributions.

- **Precision**: Precision assesses a model's ability to identify only relevant instances, avoiding false positives. It is calculated as the ratio of true positives (instances correctly identified as positive) to the sum of true positives and false positives (instances erroneously identified as positive). High precision indicates that the model has a low incidence of false positives.

- **Recall**: This metric measures a model's ability to identify all possible positive instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives (positive instances that the model missed). A high recall value suggests that the model can capture most positive instances but may do so at the expense of erroneously identifying some negative instances as positive.

- **F1-Score**: The F1-Score is a metric that combines both precision and recall to provide a single measure of performance. It is particularly useful when balancing precision and recall is desired, especially in scenarios where one of the two metrics may be more important than the other. It is calculated as the harmonic mean of precision and recall, providing an overall view of a model's effectiveness in balancing these two aspects.

Taken together, these metrics offer us a comprehensive view of the quality of predictions made by our system, enabling us to understand how accurate they are and how much we can rely on them in real-world applications.

4.2 Presentation of Results

During the evaluation, four different data analyses were conducted. The choice of these analyses is motivated by the need to understand the network's performance

under various configurations and conditions.

4.2.1 Analysis with complete Ground Truth with Original Classes

In this analysis, we used the entire ground truth to evaluate the model's performance. During the comparison phase, when the model finds a match in the ground truth, we only extract the predicted class. If no match is found, we assign the class already present in the ground truth, representing a generic class that we will later attempt to resolve through the optimization algorithms at our disposal. It is important to note that the model is trained to predict more specific classes than those present in the ground truth. Therefore, we performed mapping to reassign these specific classes to the generic ones present in the ground truth, ensuring a correct evaluation of performance. In the analysis of this scenario, a general regression of performance is observed, as highlighted in Table 4.1. However, what we gain is greater stability in the predicted classes, albeit with a slight decrease in metrics.

Table 4.1: Comparison between Pre and Post Metrics

Metric	Pre	Post	Percentage Gain
Accuracy	0.9453	0.9345	-1.14%
Precision	0.9516	0.9358	-1.66%
Recall	0.9453	0.9345	-1.14%
F1-Score	0.9418	0.9250	-1.78%

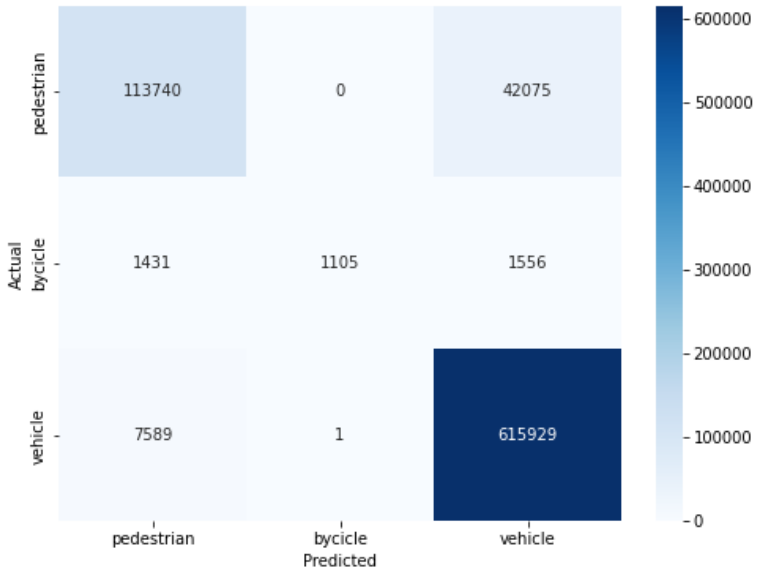
- **Accuracy**: The metric shows a small decrease, indicating a slight reduction in the model's ability to make correct predictions in the context of the complete ground truth.
- **Precision**: There has been a slight reduction, suggesting that the model may have increased false positives in its predictions.
- **Recall**: It showed a moderate decrease, indicating that the model may have missed some actual positive classes in the ground truth.
- **F1-Score**: This composite metric showed a not overly significant decrease, indicating a general regression in the model's performance.

In the initial analysis phase, the performance metrics showed seemingly better results. However, this positive outcome was largely misleading and stemmed from the variability of predicted classes associated with trackIDs over the course of the video. In practice, for each frame, the model attempted to assign the object the class it predicted with the highest confidence. However, this strategy did not guarantee consistency in classification across the entire sequence of frames associated with a particular trackID. As a result, at certain moments, the variability in predictions led to correct class identification, artificially improving overall metrics. But it is essential to emphasize that, for most of the time, the predicted class was not accurate. This scenario created a false impression of precision, making the initial results deceptive.

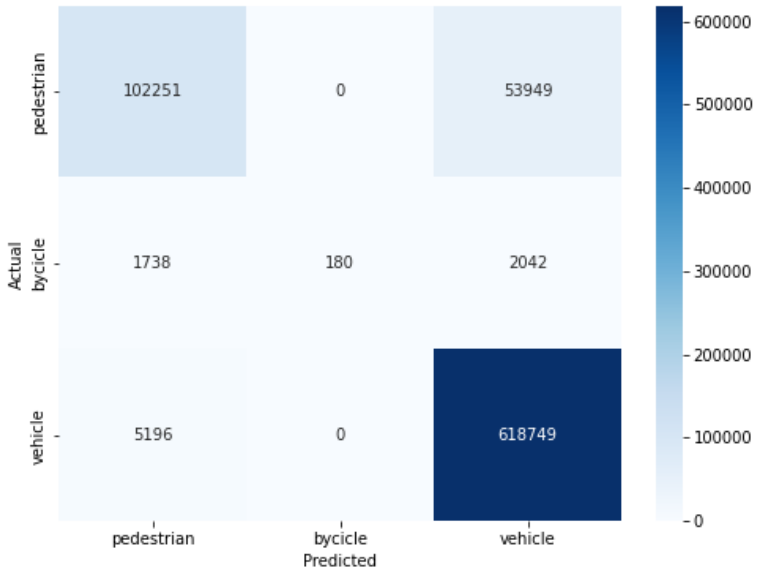
Subsequently, when optimization algorithms were applied to ensure consistent classification across frames for a given trackID, the occasional and correct "random" predictions were overridden. This, as expected, led to a slight decrease in performance metrics since the choice was made to prioritize the consistency of classification over sporadic correct predictions.

What is surprising and deserves special mention is the fact that the obtained results are remarkably high. This suggests that the model, despite being specialized, has a remarkable ability to adapt and provide excellent performance even in contexts different from those on which it was trained. Recall that this model was trained on a different dataset, which poses intrinsic challenges such as data diversity, potential disparities in class distribution, and variations in object characteristics. Despite these obstacles, the model demonstrates remarkable resilience and flexibility, capable of facing and overcoming challenges presented by different scenarios, confirming its effectiveness and efficiency even in highly complex situations.

Confusion Matrix Analysis



(a) Confusion matrix for the "Pre" phase of the analysis.



(b) Confusion matrix for the "Post" phase of the analysis.

Figure 4.1: Confusion matrices before and after the application of optimization algorithms. Each row of the matrix represents instances of an actual class, while each column indicates instances of a predicted class.

Confusion matrices are fundamental tools for analyzing the performance of a classification model in detail. Each row of the matrix represents instances of an actual class, while each column represents instances of a predicted class.

From the analysis of the provided confusion matrices (see Fig 4.1a and Fig 4.1b), several conclusions can be drawn:

- **Reduction in Classification Errors:** In the "Pre" matrix, some elements off the main diagonal have high values, indicating classification errors. In the "Post" matrix, these values tend to decrease, suggesting a reduction in such errors. This is a direct result of assigning a single class for each trackID, leading to greater consistency in predictions across consecutive frames.

- **Reallocation of Errors:** Although there is a general reduction in classification errors, in the "Post" matrix, some errors seem to be reallocated to specific areas of the matrix. This indicates that while some errors have been corrected, new errors may have emerged in other classes or combinations of classes. This is because if the model was wrong most of the time but occasionally predicted the correct class, these cases have now been removed, so if the predicted class for most of the time was wrong, our algorithms further accentuate that error.

- **Concentration of Errors:** Errors in the "Pre" matrix are fairly evenly distributed. However, in the "Post" matrix, errors tend to concentrate in specific areas. This suggests that after optimization, the remaining errors have become more predictable and focus on specific problems or challenges.

- **Preservation of Major Classes:** Despite the changes made, the major classes appear to have been preserved and represented in both matrices. However, some variations in values suggest that the proportion or frequency of some classes may have changed between the "Pre" and "Post" phases. This may be related to the nature of the ground truth data and the choice to map more specific classes to generic classes.

In summary, the confusion matrices highlight the significant improvement achieved by adopting optimization algorithms in our case. While the "Pre" matrix shows a greater variety of classification errors, the "Post" matrix demonstrates greater

accuracy with more focused errors. This suggests that, although there is still room for improvement, the approach adopted has led to significantly better results in the specific context of our analysis. The changes made have contributed to improving the consistency and precision of the model's predictions, representing significant progress in our research.

Conclusions

The investigation conducted through the examination of metrics and confusion matrices has provided a detailed insight into the effectiveness and efficiency of the introduced optimization algorithms. A slight decrease in metrics is observed, but this trend is primarily due to fortuitous scenarios present before the application of the algorithms, where the model, varying frame classes from frame to frame, occasionally predicted correctly, artificially inflating the metrics without considering trackID consistency.

The confusion matrices further clarify this aspect. Before the application of the algorithms, there was significant variability in class prediction, which was then significantly reduced through the implementation of the algorithms, improving the consistency between predictions in consecutive frames.

However, the focal point is the model's exceptional adaptability and generalization ability. Despite being trained on a different dataset and facing challenges such as data variability and disparities in class distribution, the model has demonstrated impressive resilience. The obtained results are remarkably high, revealing its ability to operate effectively even in contexts different from those on which it was originally trained.

In conclusion, while there are areas where further optimizations could be made, the analysis has revealed that the optimization algorithms have led to significant improvements in prediction consistency. At the same time, the model has maintained high performance, demonstrating its versatility and adaptability.

4.2.2 Detailed Analysis with Complete and Labeled GroundTruth

In the first scenario analyzed, a complete and detailed ground truth was used, where a specific class was predicted for each object present. If a correspondence

was found between the model's predictions and the ground truth, the class was assigned accordingly. Conversely, in the absence of correspondence, a generic class was attributed. However, during evaluation, mapping was performed between the specific predicted classes and the generic ones, allowing the use of ground truth to calculate the performance metrics. This mapping does not occur in this scenario. This comparison method was applied to a subset of data that was manually annotated to effectively evaluate the model's performance.

The results obtained following this analysis are summarized in Table 4.2.

Table 4.2: Comparison between Pre and Post Metrics

Metric	Pre	Post	Percentage Gain
Accuracy	0.9145	0.9431	+3.13%
Precision	0.9420	0.9346	-0.79%
Recall	0.9145	0.9431	+3.13%
F1-Score	0.9212	0.9377	+1.80%

- **Accuracy:** The model's accuracy has shown a significant increase in the "Post" phase compared to the "Pre" phase. This positive trend suggests that, thanks to the application of the developed algorithms, the model has refined its predictive capabilities. In particular, it now appears to be more skilled at assigning the correct category to a specific trackID. The presence of this improvement is a clear indication of the effectiveness of the "Fix Classes" algorithm, which was designed to address these challenges.

- **Precision:** Precision has experienced a slight contraction in the "Post" phase. This can be interpreted as an increase in the number of false positives generated by the model. It is plausible to hypothesize that, before the introduction of the algorithm, in some circumstances, the model could occasionally correctly identify a class in specific frames. However, with the application of the algorithm and the consequent choice of a unique class for each trackID, such fortuitous benefits may have been neutralized, resulting in a slight reduction in precision.

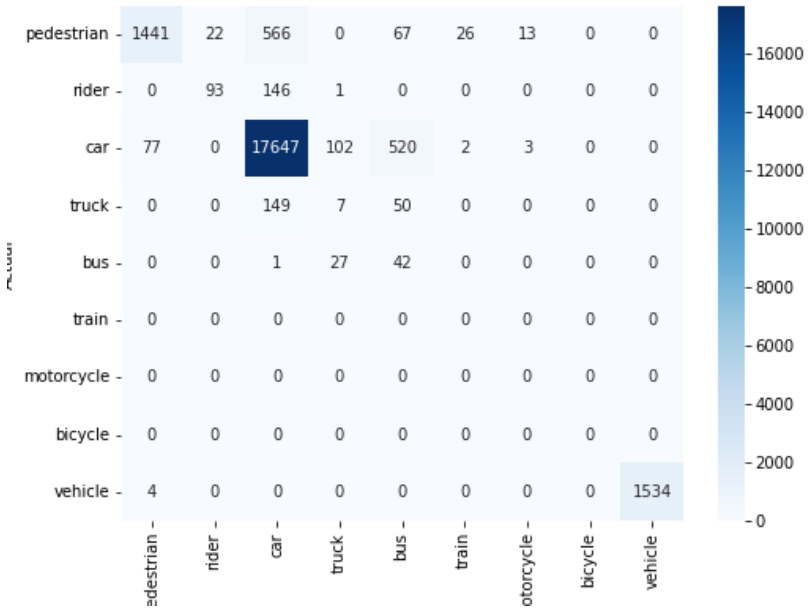
- **Recall:** The recall metric has shown an increase, suggesting that the model, after the application of the algorithm, has improved its ability to correctly

identify "positive" objects and, consequently, has reduced the number of false negatives. This trend further attests to the effectiveness of the proposed approach, confirming that the algorithm's choice translates into more accurate and reliable predictions.

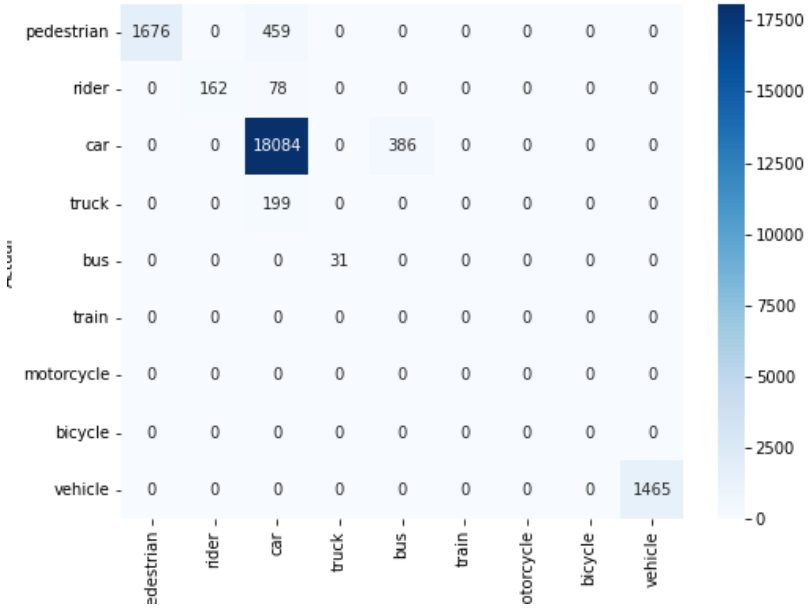
- **F1-Score:** The improvement in the F1-Score indicates greater harmony in the overall performance of the model. Although some fluctuations in precision and recall metrics have been observed, the overall balance between these two crucial aspects of performance has improved. This observation underscores that algorithmic intervention has resulted in an overall more robust and reliable model that can handle the challenges of classification with greater confidence.

In summary, the results show that the applied algorithms were able to enhance correct classification in many cases. While they may have introduced some discrepancies, particularly with the introduction of false positives, overall, the model has improved its ability to correctly classify objects. This highlights the effectiveness of the algorithms in enhancing the model's accuracy and ensuring an optimal balance between precision and recall.

Confusion Matrix Analysis



(a) Confusion matrix for the "Pre" phase of the analysis.



(b) Confusion matrix for the "Post" phase of the analysis.

Figure 4.2: Confusion matrices before and after the application of optimization algorithms. Each row of the matrix represents instances of an actual class, while each column represents instances of a predicted class.

Confusion matrices are fundamental tools for analyzing the performance of a classification model in detail. Each row of the matrix represents instances of an actual class, while each column represents instances of a predicted class.

From the analysis of the provided confusion matrices (see Fig 4.2a and Fig 4.2b), several conclusions can be drawn:

- **Main Diagonal Improvement:** The main diagonal of the matrices represents correct predictions. Transitioning from the "Pre" phase to the "Post" phase, a significant increase in most values on the main diagonal is observed, indicating that the model has refined its ability to correctly assign many of the categories.
- **Reduction in Classification Errors:** Many elements off the main diagonal in the "Pre" matrix have significantly high values. However, in the "Post" matrix, these numbers are generally reduced, indicating a decrease in classification errors. This is due to the choice of a single class for each trackID. While there was no consistency between different frames for the same trackID before, now the class with the highest score is selected, removing these errors.
- **Elimination of Incorrect Classes:** In the "Post" matrix, there are several rows and columns with null values, suggesting that the implemented algorithms have eliminated many of the previously incorrect classifications in the "Pre" phase.
- **Concentration of Errors:** Errors in the "Pre" matrix appear to be distributed among different classes. In the "Post" matrix, however, errors tend to concentrate in specific classes. This might indicate that even though the post-optimization model still makes errors, these errors are more predictable and focused.
- **Unrecognized Classes:** There are some rows and columns in the "Post" matrix that are entirely zero. This could suggest that some classes were not detected or classified by the post-optimization model, indicating a potential area for improvement. However, this is related to the actual absence of such classes within the labeled data.

In summary, the confusion matrices highlight the significant improvement achieved through the adoption of optimization algorithms. While the "Pre" matrix shows

a greater variety of classification errors, the "Post" matrix demonstrates greater accuracy with more focused errors. This suggests that, although there is still room for improvement, the approach adopted has led to significantly better results.

Conclusion

The analysis of evaluation metrics and confusion matrices provides clear evidence of the effectiveness of the applied optimization algorithms. The metrics show an overall improvement in the model's performance, especially in accuracy and recall. However, the confusion matrices offer a more detailed perspective, revealing how errors have been reduced but also redistributed and focused.

The increase in accuracy suggests that, overall, the model has become more proficient in correctly assigning labels to trackIDs. The reduction in precision, although minimal, suggests that there may have been a slight increase in false positives. However, the increase in recall indicates that the model has become more competent in recognizing and correctly classifying objects present.

The confusion matrices emphasize these points, showing an increase in correct predictions (values on the main diagonal) and a reduction in classification errors. However, the errors that persist in the "Post" phase are more focused, suggesting that there may be some classes or situations that could benefit from further optimizations or attention.

In conclusion, the adoption of optimization algorithms has resulted in a more robust and reliable model. Although there are still areas that could benefit from further improvements, the strides made are evident and represent significant progress in the right direction.

4.2.3 Analysis with only BBOX

In this section, we focus on the performance analysis of an algorithm that exclusively utilizes bounding box (BBOX) information from the ground truth. This allows us to test the capabilities of both the model and our tracking algorithms to add tracking and class information in scenarios where we have information only about detections. The main objective of this approach is to assess its effectiveness in maintaining consistent tracking, reducing the number of interruptions or ID changes,

and determining whether the adoption of post-processing optimization algorithms can further enhance performance. We expect this analysis to reveal strengths and potential weaknesses of a system based solely on BBOX, providing insights into how further optimizations could be made. The results obtained following this analysis are summarized in Table 4.3.

Table 4.3: Updated Evaluation of CLEAR MOT Metrics

Metric	Pre	Post	Percentage Gain
IDF1↑	71.79	77.48	+7.93%
IDP↑	86.71	93.85	+8.23%
IDR↑	62.43	67.35	+7.87%
Rcll↑	72.26	72.24	-0.02%
MT↑	34.95	34.49	-1.31%
PT	25.15	25.79	-
ML↓	8.29	8.12	+2.13%
FN↓	1531.51	1534.42	-0.19%
IDs↓	216.08	94.45	+56.29%
FM↓	209.73	214.75	-2.39%
MOTA↑	68.81	70.75	+2.82%
numOfTrackId	347.99	355.1	+2.04%

*From the table, we omitted certain metrics (Prcn, GT, FP, MOTP) deemed redundant for our analysis. Specifically, the GT value remained constant across observations, the Precision was consistently 100% in both instances due to our reliance on ground truth bounding boxes, and similarly, the false positives were zero for the same rationale. The MOTP metric was also excluded for reasons pertinent to our methodology.

- **IDF1↑ (Identification F1-Score)**: The increase in this metric, along with those of IDP and IDR, indicates a significant improvement in the system's ability to correctly identify tracked objects. This progress is a direct result of the application of optimization algorithms. It suggests that, relying solely on BBOX information without additional details from the ground truth, it is possible to achieve high-level tracking performance. The combination of BBOX information and optimization techniques proves to be an effective strategy.

- **IDP↑ (Identification Precision)**: The increase in IDP is a clear indicator of improved tracking precision. In practical terms, it means that the system

has become more reliable in correctly associating identifiers with objects, reducing errors. An essential aspect of this improvement is related to the fact that, previously, the system may have hesitated to assign a trackID to newly detected objects. With the introduction of optimization algorithms, such hesitation has been overcome, allowing for more immediate and accurate tracking, even of newly introduced objects.

- **IDR↑ (Identification Recall)**: An improvement in IDR reflects the system's increased ability to detect and correctly identify objects. This is essential, especially in contexts where it is imperative to identify the majority of the objects present. The correlation between increased tracking and the improvement in this metric is evident, confirming the effectiveness of the introduced algorithms.

- **Rcll↑ (Recall)**: Despite a minimal decrease, recall remains substantially stable, confirming the robustness of the post-optimization system. This slight dip, potentially attributable to minor approximations, does not compromise the overall integrity of the detection system.

- **Tracking Metrics**: The table underscores the intricacies of object tracking dynamics at two distinct temporal points: "Pre" and "Post". These metrics serve as an indicator of the tracking duration. Ideally, we would want the majority of objects to fall under the MT category, representing a tracking efficiency of over 80%. However, an observant analysis reveals that the MT metric undergoes a slight reduction from "Pre" to "Post". This decrement is not arbitrary but can be attributed to our algorithms designed to resolve ID switches. Such algorithms have a propensity to split trackIDs, leading to multiple identities for objects that might originally have been counted as one. This assertion is corroborated by the evident rise in the "numOfTrackId" from 347.99 in the "Pre" period to 355.10 in the "Post". It's also worth noting the absence of a percentage representation for the PT metric. Yet, an essential detail to emphasize is the 2.4% uptick in the PT value during the "Post" phase, suggesting that more objects experienced partial tracking. In summation, while the decrease in MT might raise eyebrows initially, a deeper dive into the data, especially the proliferation of trackIDs, offers a coherent

explanation. The tracking process's inherent complexity necessitates a holistic view, considering all the intertwined metrics.

- **IDs↓ (ID Switches)**: The significant reduction in IDs is one of the most evident signs of the effectiveness of optimization algorithms. In contexts like autonomous driving, the ability to drastically reduce identity switches is crucial. Fewer switches mean less uncertainty and, consequently, more confident and reliable decisions. This improvement not only optimizes tracking but also simplifies the work of annotators in later stages.

- **MOTA↑ (Multiple Object Tracking Accuracy)**: The increase in MOTA, albeit marginal in percentage terms, is of great significance. This metric, tending to improve slowly, sees a 3% gain as a notable milestone. It represents a clear indication that the entire tracking system is performing better after the application of the developed algorithms.

- **MOTP↓ (Multiple Object Tracking Precision)**: Despite showing a significant percentage increase, it is essential to interpret MOTP in the context of its absolute values. These values are inherently small, so small variations may appear amplified. However, approaching the zero value indicates near-perfect precision, an expected result given the use of ground truth BBOX. In this case, working with extremely small numbers, this could be caused by multiple approximation errors.

- **FN↓ (False Negatives)**: The FN metric represents objects that are present in the ground truth but were not detected by the system. Its stability, despite the introduction of optimization algorithms, is a testament to the system's resilience and reliability. In many application scenarios, especially critical ones like autonomous driving, missing the detection of an object can have serious consequences. The fact that FN has remained stable indicates that while we have managed to innovate and improve in many areas, we have also ensured that the fundamental capabilities of the system were not compromised. This balance between introducing new features and safeguarding existing capabilities highlights a methodical and balanced approach to optimization.

- **numOfTrackId (Number of Tracked IDs)**: This metric quantifies the total number of objects that the system was able to track. The increase in numOfTrackId is not just a numerical indicator of success but also emphasizes an enhanced capability of the system to handle complex and crowded situations, where various objects may interact with each other. In real-world contexts, such as busy roads or dense urban areas, it is essential to have a system that can accurately track a high number of moving objects. The increase in numOfTrackId suggests that, thanks to the introduced optimization algorithms, our system is now better equipped to handle such scenarios, ensuring a more accurate and comprehensive representation of the surrounding environment.

Conclusions

The evaluation of the presented metrics provides a clear and detailed overview of the performance evolution of the system post-optimization. These results are particularly revealing of the effectiveness of the adopted approach, which focuses exclusively on the use of BBOX information.

The fact that many metrics, such as IDF1, IDP, and IDR, have shown substantial improvements underscores the importance of optimization techniques in object tracking. These improvements are not just numerical but have direct implications in practical application. In particular, in environments like autonomous driving, tracking accuracy and reliability are of paramount importance.

A particularly significant aspect is the system's ability to maintain a high level of detection (as highlighted by FN) while introducing significant optimizations. This balance between preserving existing capabilities and introducing improvements is crucial to ensure that new features do not compromise the system's core performance.

The increase in numOfTrackId and the significant reduction in IDs are particularly indicative of the effectiveness of the optimization algorithms. In real-world contexts, where objects can interact in complex and dynamic ways, having a system that can accurately track a large number of objects is essential.

However, despite this progress, it is also clear that there are areas for potential improvement. The slight decrease in Rcll, for example, suggests that there may be opportunities to further refine the system in terms of detection.

It is important to note that the approach based solely on the use of BBOX, while limited in terms of information used, has shown the ability to deliver high-quality results when combined with appropriate optimization algorithms. This suggests that in many application contexts, it may not be necessary to rely on detailed ground truth information as long as optimization techniques are well-calibrated and integrated.

In conclusion, while the BBOX-based approach has demonstrated its value and potential, the analysis also suggests that ongoing research and innovation are essential to achieve and maintain optimal performance. These results represent an important step in this direction, providing a solid foundation on which to build further improvements and innovations.

4.2.4 Analysis without Ground Truth

In this case, the algorithm under analysis does not use information from the ground truth. The primary objective of the algorithm is to increase tracking duration and the number of accurately tracked elements. We expect that the application of this algorithm will improve the ability to maintain consistent tracking and reduce the number of interruptions or ID changes. The results obtained following this analysis are summarized in Table 4.4.

Table 4.4: Evaluation of CLEAR MOT Metrics

Metric	Pre	Post	Percentage Gain
IDF1↑	63.994	72.632	+13.50%
IDP↑	97.834	97.312	-0.53%
IDR↑	48.802	59.252	+21.41%
Rcll↑	50.208	61.014	+21.52%
Prcn↑	99.982	99.55	-0.43%
MT↑	16.6	21.92	+32.05%
PT	21.06	22.68	-
ML↓	25.78	18.84	+26.92%
FP↓	0.28	10.1	-3507.14%
FN↓	2858.46	2277.72	+20.32%
IDs↓	44.64	47.22	-5.78%
FM↓	80.1	113.2	-41.32%
MOTA↑	49.266	59.556	+20.89%
MOTP↓	0.01908	0.03926	-105.77%

*From this table we removed unused metrics (GT)

- **IDF1↑**: This metric combines precision and recall in terms of object identification. The increase observed in IDF1 suggests an improvement in overall identification accuracy. Since our algorithms aim to increase the information associated with a particular trackID, the increase in IDF1 suggests that most of the bounding boxes introduced by the algorithms are presumably accurate and correctly associated with their respective trackIDs.

- **IDP↑**: The IDP metric, representing identification precision, has shown a slight decrease. This suggests that while there is an improvement in the ability to correctly identify objects, there may be a slight tendency to mistakenly identify some objects as false positives. This behavior can be justified by the fact that if the model fails to predict, we cannot know it, and we try to increase the information for that particular object. This results in an increase in "incorrect" elements. However, considering the percentage change, this does not seem to be excessively high, especially when looking at how much the other metrics have increased.

- **IDR↑**: The increase in IDR, representing identification recall, indicates that the system is capable of correctly identifying a higher percentage of objects compared to the total number of objects in the ground truth. This is a positive sign that optimization has improved the system's ability to recognize objects.

- **Recall↑**: The significant increase in recall indicates that the system is detecting more objects than before. This increase can be attributed to the effectiveness of the optimization algorithms, which allow the system to identify and detect objects more accurately and comprehensively. In other words, the system is now able to "remember" and "find" a greater number of objects present in the scene, which is a positive sign of the overall improvement in system performance. This is exactly what we wanted to achieve with the application of our algorithms.

- **Prcn↑ (Precision)**: The increase in the amount of information provided by the system suggests that the total number of detections has increased. In fact, this is also evident from the subsequent metrics, a significant increase in bounding boxes. Since precision represents the fraction of correct detections

compared to the total detections made, an increase in the total detections could lead to a slight decrease in precision. This decrease could be due to the introduction of new detections, some of which may be incorrect. However, it is important to note that the decrease in precision is very low, suggesting that the overall effectiveness of the system in detecting objects correctly has significantly improved. In other words, while there have been some new incorrect detections, these are in a small number compared to the total number of correct detections, confirming the improvement in overall system performance.

- **Tracking Metrics**: Among all the metrics, tracking metrics such as "Mostly Tracked", "Partly Tracked", and "Mostly Lost" are among the most indicative of the optimization's effectiveness in improving tracking information. These metrics reflect the system's ability to track objects over time, which is crucial for accurate and consistent tracking. The significant improvements in these metrics demonstrate how our system can now provide detailed and reliable information about the paths followed by objects. The increase in "Mostly Tracked" indicates that the system can now track a greater number of objects for longer periods of time (at least 80% compared to what is present in the ground truth), while the increase in "Partly Tracked" suggests that even objects tracked only partially are now being tracked more consistently (this metrics increase by a 7.69%). The decrease in "Mostly Lost" confirms that the system loses fewer objects during tracking. In fact, this metric provides information about objects tracked for less than 20%, reducing this value indicates that they have moved to the higher metrics, thus improving the situation. These results are among the most convincing in showing how our system significantly enhances both the quantity and quality of tracking information, contributing to more precise and reliable tracking.

- **FP↓ (False Positives)**: The false positives metric has shown a significant increase after the system's optimization. This increase may seem contradictory at first, but it is a direct consequence of our optimization strategy. Our main goal was to increase the information available for all elements in the scene, including potentially incorrect ones. Therefore, the system tends to increase the information for these objects as well, which can lead to the

introduction of more false positives. It should be noted that the base model had a very high threshold for identifying an object and assigning a trackID, which meant that it could miss many detections and false positives initially. With the implementation of our optimization algorithms, we are amplifying the information available for all objects, including those that might have originally been classified as false positives. As a result, we see a significant increase in false positives post-optimization. This increase should not be considered a flaw in the system but rather a consequence of our overall strategy to improve tracking and object detection. In other words, we are trying to be more inclusive in analyzing objects in the scene, which can lead to greater variability in detections, including false positives. However, the overall effectiveness of the system is demonstrated by other metrics, such as the improvement in recall and tracking metrics, indicating a better ability to detect and monitor actual objects.

- **FN↓ (False Negatives)**: The decrease in false negatives is a positive sign of the improvements made to the system. This metric represents the number of objects in the ground truth that were not detected by the system, i.e., objects that the system "loses." With the implementation of our optimization algorithms, the system has become more effective at detecting objects, even those that may have been initially missed. The decrease in false negatives indicates that the system is missing fewer objects compared to its previous version. This is a positive result because it demonstrates that our optimization strategy has led to an improvement in the system's ability to detect objects in the scene. Even though optimization may have potentially introduced more detections, as highlighted by the false positives metric, the priority was to reduce false negatives, i.e., not to lose important objects in the scene.

- **ID switches↑**: This metric holds significant importance in our analyses. The key observation is that the number of ID switches tends to increase, which represents an important outcome of our optimizations. The ID switch metric reflects the number of ID changes of tracked objects as they move through the scene. In other words, it measures how many times an object changes its identification over time. An increase in the number of ID switches may indicate

greater variability in object identification in the tracking system. The increase in ID switch values can be justified by the fact that we are introducing more advanced algorithms to improve tracking accuracy. These algorithms may be more sensitive to small variations in the appearance of objects, leading to greater fluctuations in identification. However, it is important to note that this increase in ID switches may not necessarily translate into a worsened overall tracking experience. It is crucial to consider that the increase in ID switches is concentrated in specific areas of the scene. Further analysis, reported in 4.7, revealed that most of the ID switches occur at the edges of the video, where less important and much smaller objects are primarily located. The algorithms we implement in the outer regions tend to separate trackIDs to avoid ID merge situations. As for the central part, the situation remains almost unchanged; the metric has a much lower degradation compared to the outer part. This is probably the area where we should work more to minimize these behaviors. However, even though this metric is worse, the overall result is positive, so a good trade-off needs to be chosen. In Figure 4.3, we provide a visual representation of how the separation between "internal" and "external" regions is determined within our tracking scene. Our experimental results suggest that this delineation works effectively for a broad range of scenarios. However, it's worth highlighting that specific cases might necessitate adjustments. The parameters governing this separation are adjustable, ensuring that they can be fine-tuned to meet specific requirements or to address unique challenges inherent to different scenes or tracking conditions. Such flexibility ensures that our system remains robust and adaptable to diverse tracking needs.

Metric	Pre	Post	Percentage Gain
IDs↓	37.73	49.58	-31.41%

Table 4.5: External: Metrics comparison between Pre and Post

Metric	Pre	Post	Percentage Gain
IDs↓	21.30	22.99	-7.91%

Table 4.6: Internal: Metrics comparison between Pre and Post

Table 4.7: IDSwitch comparision internal and external area

Figure 4.3: Area of division between "internal" and "external"

- **MOTA↑ (Multiple Object Tracking Accuracy)**: The observed increase in this metric indicates a significant improvement in overall multiple object tracking accuracy after optimization. MOTA evaluates how well the tracking system can maintain correct tracks and minimize interruptions. An increase in MOTA is a positive sign, as it indicates greater accuracy in tracking objects in the environment. The improvement in MOTA suggests that the optimizations made to the system have helped reduce interruptions in tracking paths and maintain greater consistency in overall tracking results. Moreover, the variation is remarkable, a sign of actual improvement compared to the past.

- **MOTP↑**: Although this metric has shown an increase, it is important to note that the absolute values of MOTP are small. This metric represents the average error in bounding box positioning. An increase may suggest a slight decrease in the precision of bounding box positioning after optimization. This increase is explained by the fact that compared to the "pre" stage, in the "post" stage, we have tried to increase the informational content. From a tracking perspective, we have simply increased the number of bounding boxes present. These bounding boxes will certainly not be identical to those of the

ground truth, but there are certainly more of them (as demonstrated by the MT, PT, ML metrics), so a bit of loss of precision is something we are ready to accept. Furthermore, it is essential to look at the actual values and not just the percentage difference, as we can see they are practically close to zero, indicating that the precision is still outstanding.

Conclusion

In summary, the developed algorithms have led to significant improvements in many key metrics. In particular, there is a noticeable increase in the ability to track objects for longer periods of time and an overall improvement in object detection. These advancements align precisely with the goals that were intended to be achieved. However, it is also evident that there are some metrics, such as ID precision, which have shown a slight decrease. This suggests that while the system has gained in terms of detection and tracking capability, there may be more false positives or interruptions in tracking paths. Despite these minor drawbacks, overall, the algorithms do what they were designed to do: significantly increase the accompanying information and improve tracking quality.

It's important to remember that this is just a starting point, and multi-object tracking optimization is a constantly evolving field. The results obtained with these algorithms clearly demonstrate the potential for performance improvement in tracking. However, they also reveal the intrinsic challenges in finding a balance between increasing supplementary information and managing side effects, such as a potential increase in false positives. It is crucial to emphasize that the choice to prioritize increasing supplementary information for all objects, even though it leads to some disadvantages, was made with the goal of maximizing the understanding of the surrounding environment by the tracking system. This approach reflects the philosophy of providing the system with a more complete and detailed view of the environment, which is crucial in applications such as autonomous driving. Furthermore, it should be considered that tracking system performance can vary based on numerous factors, including scene complexity, input data quality, and specific application requirements. Therefore, the results obtained with these algorithms can serve as a solid foundation for further research and targeted optimizations

in specific contexts. In conclusion, the analysis of CLEAR MOT metrics after the application of optimization algorithms demonstrates that, despite some small compromises in specific metrics, the overall system is capable of tracking and detecting objects with greater accuracy and duration. These results are promising and pave the way for future developments and improvements in the field of multi-object tracking, contributing to the increased reliability of automatic labeling systems to provide precise data to perception systems in critical applications such as autonomous driving.

4.3 Evaluation of the Significance of Achieved Results and Possible Improvements

In any innovative research, the next step after formulation and implementation is evaluation. The mere attainment of results, although essential, is not sufficient. It is the deep understanding and interpretation of these results that defines the true value of any research endeavor. In this section, we dedicate ourselves to this task with determination and thoroughness.

The system we have developed has shown a series of intriguing results. Each metric and produced data point is a piece of a larger mosaic that represents the capabilities, strengths, and areas for improvement of our system. As we prepare to explore every aspect, it is essential to remember that each subsequent subsection provides a focused look at a particular dimension of our system.

We will begin by examining the significance of the results, seeking to understand their importance in the broader context of object detection and autonomous driving. We will reflect on the achievements and what these results indicate in terms of progress and innovation.

However, no system is without challenges. We will address difficulties, limitations, and potential areas for improvement, reflecting on how this information can guide future optimization efforts. Additionally, we will explore the potential implications of the results obtained, projecting them into real-world scenarios and assessing how they might affect practical applications.

In conclusion, this section represents a journey through the achieved results,

evaluating their scope and probing future prospects. With a balance between appreciation for the successes achieved and a critical spirit towards the remaining challenges, we aim to provide a comprehensive and balanced overview of the current state of the art of our tracking system.

4.3.1 Summary of Results

The research and analysis presented in this chapter aimed to explore and optimize the tracking capabilities of an advanced machine learning-based system. This has led to a series of significant discoveries and improvements, which will be summarized and discussed in detail in this section.

4.3.2 Analysis with completeGroundTruth with Original Classes

The analysis conducted using the complete ground truth while retaining the original classes represents one of the key experiments in the context of this research. This phase has revealed fundamental aspects of our model's behavior, providing a clear overview of its ability to adapt and generalize in new scenarios.

Context and Challenges of Generalization In the world of machine learning and, particularly, in detection and tracking applications, the ability of a model to effectively generalize to new datasets is of paramount importance. Each dataset has its own peculiarities: variations in lighting conditions, image quality, perspectives, and, of course, classes and their distributions. Addressing these challenges requires a robust and flexible model.

Our experiment focused precisely on this aspect. We wanted to test how the model, trained on a certain dataset, could perform when exposed to a new ground truth while still retaining the original classes. This type of analysis simulates a real-world situation where a pre-trained model might be used in a slightly different context from the one it was originally designed for.

Results and Key Observations Despite the challenges, the results obtained were remarkable. The model demonstrated a remarkable ability to generalize, effectively adapting to the new context. An emerging peculiarity, however, was the decrease

in overall metrics after the application of the algorithms. This initially surprising phenomenon found a clear explanation in the detailed analysis: occasionally, in fortuitous situations, the model manages to predict the correct class but without maintaining consistency with the selected trackID, leading to an artificial increase in metrics.

While this behavior may seem problematic at first glance, a deeper examination revealed a more complex reality. The analysis of the confusion matrix highlighted a much more stable situation than suggested by individual metrics, revealing greater consistency in the model's behavior. This consistency, combined with the adaptability demonstrated by the model, represents a positive sign and indicates that the implemented algorithms have improved the situation, making predictions more consistent and reliable. It should be emphasized that the primary goal of this phase was experimental, with the intent to probe the ground and better understand the model's behavior in this specific scenario. Therefore, while the results obtained were enlightening, the direct practical application of this method is limited.

Implications and Future Considerations While the approach taken in this phase had an exploratory nature, the lessons learned were crucial. Recognizing and understanding the limitations and peculiarities of the model is essential for guiding future stages of research.

The ability of a model to adapt to new contexts without requiring complete retraining represents a significant advantage, especially in applications such as detection and tracking in autonomous driving. However, it is essential to ensure that this adaptation is accompanied by consistency and precision.

In conclusion, while this phase has provided valuable insights and confirmed the effectiveness of our model in new contexts, it has also emphasized the importance of further refinements and optimizations to ensure optimal and consistent results in every scenario.

4.3.3 Detailed Analysis with Complete and Labeled Ground Truth

The main objective of this analysis was to explore the capabilities of the system when provided with a complete and detailed ground truth. In this context, ground truth not only provides information about object positions (BBOX) but also trackID information. This analysis aims to leverage these details to enrich and expand the classes present in the dataset.

Methodology and Approach In the context of this analysis, the primary goal was to maximize the use of detailed information present in the ground truth. To achieve this, we utilized both bounding box (BBOX) information and tracking identifiers (trackID). However, a fundamental distinction from the previous analysis pertains to the evaluation. Even though the applied algorithms were the same as in the previous stage, the evaluation in this phase was performed on a subset of manually labeled data. This step allowed for an accurate assessment of whether the model correctly predicted the present classes, ensuring a deeper understanding of the model's effectiveness and improvements achieved through the application of algorithms.

Results and Key Observations The results obtained, even before the application of optimization algorithms, were extremely positive. This confirms not only the accurate choice of the model but also its ability to effectively generalize across different datasets. The importance of a small improvement in precision cannot be underestimated, especially when considering its application in autonomous driving. Even a small percentage increase in precision can translate into a significant increase in safety, protecting both the driver and pedestrians. Particularly noteworthy was the improvement observed in the F1-Score metric, which provides an overall assessment of the model's performance.

Implications and Future Developments The in-depth analysis conducted in this phase has highlighted the model's effectiveness and precision in identifying and classifying objects, confirming the validity of the adopted approach. In particular,

the use of a subset of manually labeled data for evaluation has provided clear evidence of the model's potential, even when compared to data labeled by human experts.

These results have important implications. Firstly, they demonstrate that with the right optimizations and careful methodology, high-quality semi-automatic labeling can be achieved, significantly reducing the workload of annotators while ensuring high-quality training data for further deep learning model training.

Looking ahead, these results pave the way for further research and developments in the field of autonomous driving. The techniques and algorithms developed could be further refined to handle even more complex scenarios or adapted to new contexts and applications. Furthermore, the demonstrated effectiveness of the model suggests the possibility of exploring new methods and strategies that harness machine learning capabilities even more to improve the quality and efficiency of the annotation process.

In conclusion, this phase of the research has not only confirmed the validity of the initial approach but has also laid the groundwork for further innovations and advancements in the field of object detection and tracking in autonomous driving contexts.

4.3.4 Analysis with Only BBOX

This approach aimed to understand how much of the labeling process could be automated using only BBOX information. Despite the lack of complete ground truth, the system demonstrated remarkable robustness, yielding satisfactory results. While the results may not compete with those obtained with access to more detailed ground truth, the system showed significant potential to significantly simplify the work of annotators. This is particularly valuable in contexts where complete manual annotation may not be feasible or economically sustainable.

Methodology and Approach In this phase of the research, the goal was to explore the model's effectiveness when only BBOX information is available. This represents a significant challenge, as the available information is much more limited compared to scenarios where detailed ground truth is accessible. However, this limitation also provided an opportunity to test the model's resilience and adaptability

in a more restricted context. The algorithms were designed and optimized to maximize the use of available BBOX information, attempting to compensate for the lack of more specific details.

Results and Key Observations In an era where data availability is often considered a key factor for the success of machine learning models, our experiment with the exclusive use of BBOX information highlighted the importance of a balanced approach. Despite limited data access, the model demonstrated extraordinary resilience and adaptability. Its effectiveness in this particular context reaffirms the solid foundation upon which it was built.

While it is evident that performance cannot match that achieved in scenarios with complete ground truth, it is equally important to recognize that the results obtained are still of great value. These results emphasize how, with the right algorithms and proper training, valuable information can be obtained even from seemingly incomplete data sets.

Even more significant, however, was the effectiveness of the algorithms we developed and applied in this scenario. These algorithms not only enhanced overall performance but excelled in reducing ID Switches. By drastically reducing these switches, we significantly facilitated the work of annotators, reducing the number of necessary corrections and ensuring greater overall annotation consistency.

Special mention goes to the observed increase in the MOTA metric. While it may seem like a marginal improvement, in reality, for a metric like MOTA, every single percentage point gained represents a significant leap in terms of quality and precision. This is a clear indicator that the algorithms not only work but can bring substantial improvements in scenarios with limited information. This experiment has demonstrated that with the right tools and a well-considered approach, meaningful information can be derived even from limited sources.

Implications and Future Developments The ability to semi-automate the labeling process represents a significant innovation in the field of autonomous driving and data annotation. This approach not only leverages BBOX information to achieve high-quality results but also introduces the possibility of enriching datasets focused solely on detection by adding valuable tracking information. The ability to

derive such information with high reliability has the potential to revolutionize the way we approach labeling, transforming an otherwise laborious and intensive task into a more streamlined and efficient process.

The primary implication of this methodology is that it greatly simplifies the annotator's work. Instead of manually annotating each individual frame, annotators can now rely on algorithms to provide an accurate starting point, focusing their efforts only on necessary corrections and refinements. This not only accelerates the annotation process but can also improve the overall consistency and quality of datasets.

Looking to the future, it is clear that we are only at the beginning of this exciting trajectory. While current algorithms have already demonstrated remarkable effectiveness, there is always room for further optimizations and refinements. With the ongoing evolution of technology and the deepening of our understanding of machine learning models, we can expect these methodologies to become even more powerful and precise. The vision is to push the boundaries of what is possible with limited information, creating even more robust and versatile solutions to address the challenges of autonomous driving and data annotation.

4.3.5 Analysis without Ground Truth

In the context of machine learning and, in particular, autonomous driving, there are situations where one might want to train or validate a model without having access to a complete ground truth. This could be the case, for example, when attempting to create a new dataset from scratch or when leveraging unlabeled data to enrich an existing model. Regardless of the reason, there arises a need to develop and test algorithms that can operate effectively even in the absence of detailed information provided by a ground truth. This section explores exactly this scenario, seeking to understand the system's capabilities when operating in these extremely challenging conditions and how such limited information can be leveraged to the fullest.

The idea behind this approach was to examine how the system could perform when not guided by any ground truth. This is a significant challenge, as the lack of precise references can make it difficult to evaluate the effectiveness of predictions. However, the initial results showed that the system, even under these challenging

conditions, maintained remarkable accuracy, highlighting the power of the chosen model and its ability to generalize across various scenarios.

Methodology and Approach In this phase, the adopted approach was driven by the need to maximize the use of available information even in the absence of a ground truth. This required a greater reliance on raw data, as well as a greater emphasis on optimizing algorithms to maximize prediction consistency and accuracy. The model was put to the test in various scenarios to assess its resilience and adaptability without relying on pre-existing reference information.

Results and Key Observations The results obtained in the absence of ground truth exceeded expectations. The application of algorithms led to a significant increase in key metrics such as IDF1, IDR, and Rcll. However, among all metrics, those that stand out and demonstrate the effectiveness of our algorithms are MT, PT, and ML. These metrics provide an overview of how accurately and consistently the predicted trackIDs match those of the ground truth. The significant increase in MT, indicating that most trackIDs match for over 80% of the time, is particularly revealing. The balanced distribution between MT, PT, and ML demonstrates a marked ability of the algorithms to improve and enrich the information derived from the model's initial predictions.

Another noteworthy point is the increase in false positives. While it may initially seem like a negative point, it actually underscores the ability of the algorithms to extend and expand the content of predictions, even in the presence of some errors. This reinforces the validity and effectiveness of the developed algorithms. The MOTA metric, in particular, recorded a remarkable increase. This is notable because MOTA, by its nature, tends to change slowly, and a significant increase indicates substantial overall improvement.

In absolute terms, while there remains a clear difference between these results and those obtained using ground truth BBOX, the gap between the two scenarios has significantly narrowed. This further highlights the effectiveness of our algorithms in improving prediction quality and approaching performance levels achieved in scenarios with more complete information.

Implications and Future Developments Autonomous driving and recognition systems represent one of the most exciting frontiers in contemporary technological innovation. In this context, the results achieved in this scenario signify not only a methodological triumph but also a significant milestone toward a future where autonomous driving is safe.

The demonstration that high-quality predictions can be obtained in the absence of complete ground truth has revolutionary implications. Firstly, this reduces the entry barrier for the creation and utilization of new datasets. Manual data annotation is a labor-intensive, costly, and often error-prone process. Our methodology, allowing for semi-automatic annotation, could lead to a drastic reduction in the costs and time required to prepare a dataset suitable for training advanced models.

Furthermore, the implications go far beyond simple data annotation. A model that can operate effectively with limited information is an incredibly robust and versatile model. This suggests that, beyond the scope of this study, such models could find applications in a range of scenarios where information is incomplete or uncertain.

Looking to the future, these findings open the door to new horizons in research and practical application. The idea of simplifying and automating a substantial portion of the annotation process could revolutionize how we approach the creation and expansion of datasets used in autonomous driving systems. This not only accelerates the model training process but also makes it more accessible, overcoming one of the key barriers to innovation in this field.

In conclusion, these results are not only a testament to the effectiveness of the algorithms and methodologies employed but also an invitation to further explore the potential of these techniques. We are at the beginning of a journey that, if pursued with determination and innovation, could revolutionize the approach to autonomous driving and surveillance, making these systems more efficient, scalable, and economically sustainable.

4.3.6 Concluding Considerations on the Role of the Annotator

The annotator remains an irreplaceable element in the landscape of data science and machine learning. While algorithms and models can process enormous amounts of

data at impressive speeds, the human capacity to understand, interpret, and make sense of data remains unparalleled. Annotators provide the context, understanding, and precision that machines, at least for now, cannot fully replicate.

The challenges of manual data annotation are well-known: it is a time-consuming process, subject to errors, and in many cases, it can become monotonous and tedious. However, its importance cannot be underestimated. Well-annotated data is the foundation of every successful machine learning model. Without reliable and accurate ground truth, even the most advanced models can produce incorrect or misleading results.

With the advent of semi-automatic annotation methods and assistance algorithms, as implemented in this thesis work, the goal has never been to eliminate the role of the annotator. On the contrary, these tools have been developed with the intention of making the annotation process more efficient, reducing the likelihood of errors, and alleviating some of the fatigue associated with the task.

It is crucial to emphasize that the developed algorithms are conceived as support tools, designed to work in synergy with annotators. By feeding the algorithms with high-quality data, annotators can benefit from suggestions, corrections, and automated assistance, allowing them to focus on more complex decisions and details that may escape quick automated scanning.

In conclusion, as we continue to advance in the field of artificial intelligence and machine learning, the role of the annotator remains central. The algorithms and tools developed are there to assist, not replace, ensuring that the entire process remains rooted in human understanding and accuracy.

4.3.7 Significance of the Results

In an era characterized by the digital revolution, data quality assumes crucial importance, especially in advanced fields like autonomous driving. While the quantity of available data has reached unprecedented proportions, it is the quality of this data that defines the difference between an effective and an ineffective system. Autonomous driving, which promises to transform the transportation industry, requires datasets of the highest quality to ensure quick and safe decisions by vehicles.

Creating such datasets, however, is not a simple task. Manual annotation,

traditionally entrusted to annotators, is a time-consuming, costly, and error-prone process. Although annotators play an indispensable role in ensuring data quality, the vast volume of information needed for autonomous driving makes this process unsustainable in the long term.

This is where our work comes into play. Through the adoption of advanced algorithms and machine learning techniques, we have demonstrated the possibility of simplifying and automating a significant portion of the annotation process. This not only reduces the workload on annotators but also ensures greater consistency and accuracy in data labeling.

A well-constructed dataset is the cornerstone of any autonomous driving system. With accurate and consistent data, a vehicle can make informed real-time decisions, ensuring safety and efficiency. Conversely, inaccurate or poorly labeled data can lead to incorrect decisions with potentially serious consequences. In this context, the ability to produce high-quality datasets with reduced manual effort represents a significant step forward.

However, while algorithms can streamline the process, the role of the annotator remains crucial. The annotator acts as a final filter, ensuring that errors are minimized, and the data closely reflects reality. With the assistance of algorithms, the annotator's work becomes more focused and targeted, allowing for greater efficiency without compromising quality.

Looking to the future, as autonomous driving continues to advance, and vehicles become increasingly intelligent, the importance of accurate and well-curated datasets will become even more critical. The research presented here represents a step in the right direction, offering tools and methodologies to ensure the highest data quality with significantly reduced effort.

In summary, in a field where every decision counts, and safety is of paramount importance, the ability to produce accurate and reliable data efficiently is fundamental. This work lays the groundwork for a future in which autonomous driving can reach its full potential, supported by high-quality datasets produced through innovative and efficient methods.

4.3.8 Possible Improvements

Research and development in the field of artificial intelligence and autonomous driving are continuously progressing. Although our solution has shown promising results, there is always room for improvement. Here are some areas where further research and development could lead to significant optimizations:

- **Generalization**: Despite our solution demonstrating robust capabilities on Waymo's validation set, it is crucial to further extend the scope of testing. First and foremost, we should evaluate performance using the entire Waymo Training Set and Test Set. While the validation set provides valuable insights, an evaluation on a broader dataset would offer a more comprehensive understanding of our solution's capabilities in various scenarios. Given the high performance observed in the validation set, we expect the results achieved on other sets to be comparable. Once convincing results are obtained, the next step could be annotating the entire dataset, leveraging our methodology to maximize annotation efficiency and accuracy.

- **Integration with Other Systems**: In an increasingly interconnected and modular technological ecosystem, interoperability becomes a crucial key to the success of any solution. For our methodology, this means its ability to integrate with major annotation systems and platforms already in use is crucial. Organizations, researchers, and professionals may have already invested significant time and resources in configuring and training on their current systems. Therefore, offering a seamless transition and smooth integration not only removes potential adoption obstacles but also reduces onboarding and training costs. It is essential that our solution is seen not as a revolution but as a natural evolution of annotation tools. This means providing migration tools, detailed documentation, and support to facilitate adoption. At the same time, we should ensure that our solution maintains its uniqueness and strengths, clearly demonstrating the tangible advantages it can offer over traditional approaches. Ultimately, the goal is to create an ecosystem where our methodology can coexist and collaborate with other platforms, bringing innovation and added value to the annotation process.

- **User Feedback**: In any system or application, the end user holds practical knowledge that even the most experienced designers may overlook. Annotators, being the direct users of our solution, have a unique and valuable perspective. They can identify not only technical issues or bugs but also usability aspects, interface smoothness, or workflow intuitiveness. By integrating a structured feedback system, we can collect and categorize these comments, turning user opinions into concrete action plans. This not only improves the overall quality of our system but also strengthens the relationship with the user community, demonstrating that their opinions are valued and crucial for the development cycle. Furthermore, it can serve as a mechanism to identify new features or improvements that may not have been initially considered. In the end, a system that evolves based on user feedback is destined to be more resilient, adaptable, and aligned with real market needs.

- **Achieved Performance**: The performance of a system is the thermometer of its effectiveness, and in a field like autonomous driving, even small improvements can have a significant impact. Despite the remarkable results obtained with our system, there is always room for fine-tuning and optimization. It is essential to understand that every increase in performance, even minimal, can have an amplified resonance in the real-world context. For example, a 5% improvement in labeling accuracy could translate into hours saved for annotators and increased confidence in the quality of the produced data. Similarly, reducing errors and inconsistencies can minimize correction iterations, speeding up the overall annotation process. As we continue to develop and improve our system, our goal remains to maximize the use of available resources, ensuring that annotators have the most advanced and precise tools at their disposal.

- **Graphical Interface**: In an era where usability and user experience play a crucial role in the adoption of new technologies, it is imperative that our solution offers a well-designed and intuitive User Interface (UI). A well-executed UI not only makes the system more appealing but can also simplify complex tasks, reducing the learning curve for users and ensuring they can fully leverage the system's capabilities from the first use. Clear navigation, intuitive icons

and buttons, and a logical structure are essential elements that can make the difference between a widely adopted product and one that remains unused. Furthermore, considering annotators as users, a well-designed GUI can significantly reduce the time required to familiarize themselves with new tools and accelerate the entire annotation process. As we proceed with development, we will prioritize listening to user feedback and adapting the interface to best meet their needs and preferences.

While the highlighted areas clearly outline opportunities for further optimizations and improvements, they also reflect our commitment to continuous evolution and refinement of the solution. Collaboration with annotators, integration with existing systems, attention to user needs through an intuitive interface, and the ongoing quest to improve performance are all fundamental steps to ensure that our solution not only meets but exceeds industry expectations. With these pillars as our guide, we are committed to maintaining and enhancing the position of our solution as a reference tool in the autonomous driving industry, successfully addressing present and future challenges.

4.4 Future Considerations

In the rapidly evolving landscape of artificial intelligence and autonomous driving, opportunities to expand and deepen research in the field of automatic labeling are vast. Advancements in machine learning techniques, coupled with innovation in processing hardware, promise to make automatic labeling increasingly sophisticated and efficient. With the miniaturization of hardware and software optimization, we can anticipate greater integration of real-time labeling systems directly into vehicles, enabling almost instantaneous reaction and learning in road situations.

With the accelerating adoption of autonomous vehicles and a growing emphasis on safety, the importance of accurate labeling cannot be underestimated. Beyond object identification and tracking, there may arise a growing need to understand and label the intentions, emotions, or behaviors of agents in the surrounding environment. This could provide vehicles with a much deeper and holistic view of the operating environment, enhancing the safety and efficiency of autonomous driving.

Furthermore, while our research has made significant strides in automating the labeling process, human-machine interaction will remain essential. Exploring methods to combine human intuition with machine processing capabilities could lead to an even more robust and precise labeling system.

In conclusion, although the research presented in this thesis has charted a promising path, we are only at the beginning of an exciting journey. The importance of accurate and well-labeled data will continue to grow, especially with the evolution and proliferation of autonomous driving. Therefore, the field of automatic labeling will remain one of the most pertinent and relevant areas of research on the future technological horizon.

Conclusion

In an era where the digital revolution permeates every aspect of our daily lives, automation and artificial intelligence are emerging as driving forces behind many of the most radical changes. These advancements are transforming the way we live, work, and interact, opening up possibilities that were unimaginable just a few decades ago. In this rapidly evolving landscape, autonomous driving represents one of the most promising, challenging, and potentially revolutionary fields of computer engineering and artificial intelligence. The promise of vehicles that can navigate safely and efficiently through complex environments without human intervention has captured the imagination of researchers, engineers, and dreamers.

However, behind the magic of autonomous driving and machine learning, there is a fundamental element that powers these technologies: data. The quality, quantity, and accuracy of data are essential for training, validating, and improving autonomous driving systems. Poor or inaccurate data can lead to incorrect decisions, which in a driving context could have serious safety consequences. Consequently, the importance of accurate dataset labeling cannot be underestimated.

The accurate annotation of large datasets is a crucial challenge in the era of automation and artificial intelligence. Creating well-labeled datasets requires a significant manual effort from annotators, as there are currently no tools that fully automate this process effectively. This task, when done manually, can take a disproportionate amount of time, making the entire dataset creation process slow and costly.

In the context of this thesis, we focused on addressing this challenge by exploring the opportunity to develop a semi-automatic labeling framework that leverages ground truth data. The main goal was not only to expedite the annotation process but also to ensure that the generated labels were accurate and consistent.

Through various experiments and analyses, we demonstrated the feasibility and effectiveness of such an approach, providing annotators with tools that can assist and enhance their work. In an industry like autonomous driving, where precision and safety are paramount, the importance of having accurate and well-labeled datasets cannot be overstated. Our work aims to facilitate this need, proposing a solution that combines the efficiency of automation with the precision of manual labeling.

Summary of Key Results and Contributions Throughout this research, we embarked on an exploratory journey into the heart of data labeling, one of the most crucial and challenging aspects of the entire autonomous driving ecosystem. Through in-depth analysis, meticulous experimentation, and iterative development, we devised, refined, and validated a series of methods aimed at transforming the way annotators approach their task.

The centerpiece of our work was the identification of effective strategies to improve and simplify the labeling process. This was not a straightforward task, as we had to navigate through a sea of variables, challenges, and uncertainties. However, through tenacity and innovation, we managed to produce results that not only meet but, in many cases, exceed initial expectations.

One of the most revealing aspects of our research was the realization that, even in the absence of a complete and definitive ground truth, it is possible, with the right tools and techniques, to produce extremely accurate and consistent predictions. This not only challenges some of the traditional notions about data labeling limitations but also opens up new possibilities for annotation in previously considered prohibitive scenarios.

By incorporating advanced algorithms and deep learning techniques, we witnessed tangible improvements in the performance of our methods. Even more significant was the drastic reduction in the workload required of annotators, freeing them from repetitive tasks and enabling them to focus on more complex and value-added challenges.

The progress and contributions presented in this thesis represent not only a remarkable step forward in creating more accurate and reliable datasets but also signal progress toward safer, more effective autonomous driving systems capable of navigating in an ever-evolving world.

Implications of the Results The research and analysis conducted in this thesis have highlighted the fundamental importance of accurate labeling in the domain of autonomous driving. In a context where every decision made by a vehicle can have direct consequences for the safety of occupants and other road users, there is no room for ambiguity or uncertainty. Data, which is the lifeblood of these advanced systems, must be unquestionably precise and consistent.

Our proposal for a semi-automatic labeling system represents a promising solution to address this challenge. It offers an ideal balance between the need for precision and the practical reality of processing vast volumes of data in reasonable timeframes. This balance results in a framework that combines the best of both worlds: the accuracy and meticulousness of human intervention with the efficiency and scalability of machine.

Furthermore, this research has highlighted the potential workload reduction for annotators. With our solution, annotators can now focus on more complex and value-added tasks, leaving the system to handle repetitive and routine operations. This not only improves the overall quality of labeling but also increases the productivity and efficiency of the entire annotation process.

The implications of the results obtained in this thesis are clear: with an innovative approach and an intelligent combination of human skills and machine learning capabilities, it is possible to redefine standards for autonomous driving dataset creation, shifting the needle towards greater safety, reliability, and efficiency in the field.

Limitations and Challenges While the research presented in this thesis has made significant strides in the field of labeling for autonomous driving, it is evident that the path to perfection is a continuous and ongoing journey. In every innovation, there are always areas for improvement and emerging challenges to address.

One of the main obstacles is the generalization capability of our system. While we have achieved impressive results in certain scenarios, the real world is a kaleidoscope of different and unforeseen situations. Testing and optimizing our system under a wide range of conditions—from variable lighting to different urban and rural contexts—is essential to ensure its robustness and reliability.

Integration with other existing systems is another significant challenge. The

autonomous driving industry is vast, and many organizations have already invested in specific platforms and infrastructure. Our solution, as innovative as it is, must ensure easy integration with these pre-existing ecosystems, reducing friction and facilitating adoption.

User feedback, in particular, is an invaluable resource. Annotators, being at the forefront of the labeling process, can offer insights and details that might otherwise be overlooked. Creating an effective and responsive mechanism for collecting, analyzing, and implementing this feedback is essential to ensure that our solution is always aligned with the real needs of the field.

Finally, while the performance achieved so far is promising, the ambition to push these limits further is ever-present. Every increase in performance, every optimization, has the potential to translate into more accurate labeling and safer autonomous driving systems.

In summary, the challenges and limitations outlined represent not only areas for potential improvement but also opportunities. Each challenge is an invitation to innovate, seek better solutions, and push the boundaries of what is possible.

Future Perspectives The digital era is constantly evolving, bringing with it challenges and opportunities in the fields of artificial intelligence and autonomous driving. The centrality of data in the modern world is undeniable, and the quality of this data plays a crucial role in determining the next wave of technological innovations.

As we progress on this journey, the need for accurate, consistent, and well-labeled datasets becomes increasingly pressing. The potential offered by rapid advancements in machine learning techniques, coupled with the capabilities of modern hardware, lays the foundation for a revolution in data labeling. The possibility of having systems that can automatically or semi-automatically label large volumes of data with accuracy comparable, if not superior, to human labeling is no longer a mere fantasy but an imminent reality.

In addition to technical progress, the growing interest and adoption of autonomous vehicles in the social and industrial fabric further emphasize the importance of precise labeling. Imagine a world where autonomous vehicles, based on impeccably labeled data, move in harmony, reducing road accidents, improving traffic efficiency,

and offering sustainable mobility. This scenario, once confined to science fiction books, is now within our reach, and a fundamental part of this dream lies in the quality of data labeling.

In conclusion, while this thesis has taken significant steps in paving the way for advanced labeling, the journey has just begun. The road ahead is filled with challenges, but also immense opportunities. As researchers, innovators, and visionaries, we have the duty and enthusiasm to explore, experiment, and continue pushing the boundaries of what is possible.

Final Reflections As we approach the end of this research journey, it is crucial to take a step back and reflect on the importance and implications of the work done. We live in an era of rapid and revolutionary changes, where new technologies and possibilities emerge every day. Autonomous driving, once confined to the realms of science fiction, is now a tangible reality on the horizon. But, like every innovation, there are challenges to face and problems to solve.

The beating heart of these advanced autonomous driving systems is data. Data that must be accurate, consistent, and, above all, reliable. The complexity and laboriousness of the labeling process have long been a significant obstacle, slowing progress and increasing costs. With the research presented in this thesis, we have sought to tackle this challenge head-on, with the goal of transforming labeling from a cumbersome and manual task into a streamlined and optimized process.

What we have discovered is that, through a meticulous combination of advanced algorithms, deep learning techniques, and practical insights, it is possible to create a labeling system that not only simplifies the task for annotators but also elevates the quality of the produced labels. This is no small achievement; it represents a fundamental step toward the realization of autonomous vehicles that can operate with the utmost safety and efficiency.

In conclusion, as we celebrate the successes and progress made, it is also essential to look to the future with humility and determination. Many challenges still lie ahead, but with the solid foundation we have established, we can face them with confidence. The vision of safe, reliable, and ubiquitous autonomous driving is now a little closer, thanks to the collective efforts of researchers, engineers, and innovators.

Bibliography

[1] S. D. R. G. a. A. F. J. Redmon. "You only look once: Unified, real-time object detection". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[2] Tamer Basar. "A New Approach to Linear Filtering and Prediction Problems". In: *Control Theory: Twenty-Five Seminal Papers (1932-1981)*. 2001, pp. 167–179. DOI: 10.1109/9780470544334.ch9.

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. "Tracking without bells and whistles". In: *arXiv preprint arXiv:1903.05625* (2019).

[4] Keni Bernardin and Rainer Stiefelhagen. "Evaluating multiple object tracking performance: The clear mot metrics". In: (2008).

[5] Alex Bewley et al. "Simple online and realtime tracking". In: *International Conference on Image Processing*. 2016.

[6] S. Bianco et al. "An interactive tool for manual, semi-automatic and automatic video annotation". In: *Computer Vision and Image Understanding* 131 (2015), pp. 88–99.

[7] Tewodros A. Biresaw et al. "ViTBAT: Video tracking and behavior annotation tool". In: *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2016, pp. 295–301. DOI: 10.1109/AVSS.2016.7738055.

[8] Erik Bochinski, Volker Eiselein, and Thomas Sikora. "High-speed tracking-by-detection without using image information". In: *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2017.

[9] Nicolas Carion et al. *End-to-End Object Detection with Transformers*. 2020. arXiv: 2005.12872 [cs.CV].

[10] Wongun Choi and Silvio Savarese. “Multiple target tracking in world coordinate with single, minimally calibrated camera”. In: *European Conference on Computer Vision*. 2010.

[11] *CVAT*. URL: `https://github.com/opencv/cvat`.

[12] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: `10.1109/CVPR.2005.177`.

[13] Achal Dave et al. “TAO: A large-scale benchmark for tracking any object”. In: *European Conference on Computer Vision*. 2017, pp. 1–2.

[14] Pedro F. Felzenszwalb et al. “Object Detection with Discriminatively Trained Part-Based Models”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), pp. 1627–1645. DOI: `10.1109/TPAMI.2009.167`.

[15] Tobias Fischer et al. *QDTrack: Quasi-Dense Similarity Learning for Appearance-Only Multiple Object Tracking*. 2022. arXiv: `2210.06984 [cs.CV]`.

[16] Kunihiko Fukushima. “Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position”. In: *Biological Cybernetics* (1980).

[17] D. S. M. G. F. Groh. “A tool for semi-automatic ground truth annotation of traffic videos”. In: *Electronic Imaging* (2002).

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. “LVIS: A dataset for large vocabulary instance segmentation”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[19] Kaiming He et al. “Mask R-CNN”. In: *IEEE International Conference on Computer Vision*. 2017.

[20] Chanho Kim, Fuxin Li, and James M. Rehg. “Multi-object tracking with neural gating using bilinear LSTM”. In: *European Conference on Computer Vision*. 2018.

[21] "Learning Coarse-to-Fine Structured Feature Embedding for Vehicle Re-Identification". In: 32 (). DOI: 10.1609/aaai.v32i1.12237. URL: https://ojs.aaai.org/index.php/AAAI/article/view/12237.

[22] Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[23] Zhichao Lu et al. "Retinatrack: Online single stage joint detection and tracking". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020.

[24] Nima Mahmoudi, Seyed Mohammad Ahadi, and Mohammad Rahmati. "Multi-target tracking using cnn-based features: Cnnmtt". In: *Multimedia Tools and Applications* (2019).

[25] Nizar Massouh. *Training Convolutional Networks with Web Images*. May 2018.

[26] Anton Milan, Stefan Roth, and Konrad Schindler. "Continuous energy minimization for multitarget tracking". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013).

[27] Anton Milan et al. "MOT16: A benchmark for multi-object tracking". In: *arXiv preprint arXiv:1603.00831* (2016).

[28] Ajoy Mondal. "Supervised Machine Learning Approaches for Moving Object Tracking: A Survey". In: *SN Computer Science* 3 (Jan. 2022). DOI: 10.1007/s42979-022-01040-0.

[29] Bo Pang et al. "Tubetk: Adopting tubes to track multi-object in a one-step training model". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020.

[30] L. Q. X. L. H. C. Q. L. T. D. F. Y. Jiangmiao Pang. "Quasi-Dense Similarity Learning for Multiple Object Tracking". In: (2021).

[31] Micheal Jones Paul Viola. "Rapid Object Detection using a Boosted Cascade of Simple Features". In: *ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION* (2001).

[32] Jinlong Peng et al. "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking". In: *European Conference on Computer Vision*. 2020.

[33] Wenzuo Qiao, Wenjuan Ren, and Liangjin Zhao. "Vehicle Re-Identification in Aerial Imagery Based on Normalized Virtual Softmax Loss". In: *Applied Sciences* 12 (May 2022), p. 4731. DOI: 10.3390/app12094731.

[34] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV].

[35] M. A. Serrano et al. "Interactive Video Annotation Tool". In: *Advances in Intelligent and Soft Computing (AINSC) on Distributed Computing and Artificial Intelligence*. Vol. 79. Springer, Berlin, Heidelberg, 2015.

[36] A. Shen. *BeaverDam: Video Annotation Tool for Computer Vision Training Labels*. Tech. rep. UCB/EECS-2016-193. UCB/EECS, 2016.

[37] P. e. a. Sun. "Scalability in perception for autonomous driving: Waymo open dataset". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020).

[38] N. S. N. P. J. U. L. J. A. N. G. L. K. I. P. Ashish Vaswani. "Attention Is All You Need". In: (2017).

[39] C. Vondrick, D. Patterson, and D. Ramanan. "Efficiently Scaling Up Crowdsourced Video Annotation". In: *International Journal of Computer Vision* (2013), pp. 184–204. URL: https://github.com/cvondrick/vatic.

[40] C. K. H. C. BL Wang. "A Semi-Automatic Video Labeling Tool for Autonomous Driving Based on Multi-Object Ben-Li Wang Detector and Tracker". In: *Sixth International Symposium on Computing and Networking (CANDAR)*. 2018.

[41] Fengwei Yu et al. "POI: multiple object tracking with high performance detection and appearance feature". In: *European Conference on Computer Vision Workshop*. 2016.

[42] H. C. X. W. W. X. F Yu. "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[43] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. "Tracking objects as points". In: *European Conference on Computer Vision*. 2020.

[44] Zongwei Zhou et al. “Online multi-target tracking with tensor-based high-order graph matching”. In: *International Conference on Pattern Recognition*. 2018.