electronics

MDPI

Article

Enhancing Multi Object Tracking with CLIP: A Comparative Study on DeepSORT and StrongSORT

Khadijah Alkandary *,†, Ahmet Serhat Yildiz † and Hongying Meng

Department of Electronic and Electrical Engineering, Brunel University London, London UB8 3PH, UK; ahmetserhat.yildiz@brunel.ac.uk (A.S.Y.); hongying.meng@brunel.ac.uk (H.M.)

* Correspondence: khadijah.alkandary@brunel.ac.uk

† These authors contributed equally to this work.

Abstract

Multi object tracking (MOT) is a crucial task in video analysis but is often hindered by frequent identity (ID) switches, particularly in crowded or occluded scenarios. This study explores the integration of a vision-language model, into two tracking by detection frameworks DeepSORT and StrongSORT to enhance appearance-based re-identification. YOLOv8x is employed as the base detector due to its robust localization performance, while CLIP's visual features replace the default appearance encoders, providing more discriminative and semantically rich embeddings. We evaluated the CLIP enhanced DeepSORT and StrongSORT on sequences from two challenging real world benchmarks: MOT15 and MOT16. Furthermore, we analyze the generalizability of YOLOv8x when trained on the MOT20 benchmark and applied to the chosen trackers on MOT15 and MOT16. Our findings show that both CLIP enhanced trackers substantially reduce ID switches and improve ID-based tracking metrics, with CLIP StrongSORT achieving the most consistent gains. In addition, YOLOv8x demonstrates strong generalization capabilities for unseen datasets. These results highlight the effectiveness of incorporating vision language models into MOT frameworks, particularly under visually challenging conditions.

Keywords: YOLO; DeepSORT; StrongSORT; detection; tracking; autonomous driving; CLIP; vision-language models

1. Introduction

Recent years have witnessed substantial advances in computer vision, particularly in object detection, visual representation learning, and multi-object tracking (MOT) [1–3]. These advances have been largely driven by deep learning architecture and the availability of large-scale datasets, enabling robust perception under increasingly complex and dynamic environments. Among fundamental tasks in vision, MOT remains critical due to the inherent difficulty of consistently associating object identities across video frames. This capability is essential for real-world applications such as autonomous driving [4], surveillance [5], robotics [6], and smart city infrastructure, where real-time decision making under strict computational constraints is required.

The YOLO (You Only Look Once) family of models has consistently delivered high-performance object detection with real-time inference speed. The release of YOLOv8 [7] further improves detection accuracy and inference efficiency, making it attractive for deployment in resource-constrained environments. YOLOv8 introduces architectural refinements,

Check for updates

Academic Editor: Jinsheng Xiao

Received: 29 October 2025

Revised: 23 December 2025

Accepted: 30 December 2025

Published: 7 January 2026

https://doi.org/10.3390/electronics15020265

anchor-free detection, and improved loss formulations that collectively enhance localization accuracy and generalization performance. However, while YOLOv8 performs well in per-frame detection, it lacks explicit mechanisms for maintaining identity associations over time, a core requirement in MOT tasks.

YOLOv8 was selected for this study based on a prior comparative evaluation conducted by the authors, which demonstrated that YOLOv8 provides a more stable and well-balanced trade-off between accuracy, inference speed, and generalization compared to newer YOLO variants [8]. Such stability is particularly important for MOT pipelines, where detection noise can propagate and negatively affect downstream data association. In addition, recent studies have confirmed the effectiveness of YOLOv8 within MOT frameworks, reporting strong performance across diverse benchmarks and scenarios [9–11].

To fully leverage YOLOv8's strengths in object localization, robust appearance embedding models are required to support reliable object re-identification. CLIP (Contrastive Language–Image Pretraining) [12], a large-scale vision–language model, has demonstrated a strong capability for producing semantically rich and highly discriminative visual embeddings. Unlike traditional ReID models that primarily rely on low-level appearance cues such as color, texture, and shape, CLIP learns from large-scale image–text pairs, enabling it to encode higher-level semantic attributes. These attributes include clothing style, semantic context, and object–human relationships, which are difficult to capture using conventional CNN-based ReID architectures. As a result, CLIP-based features exhibit improved robustness to occlusion, illumination changes, and viewpoint variations, and maintain stronger identity consistency across challenging frames. Although originally developed for zero-shot classification and image–text retrieval, CLIP's embeddings have recently demonstrated strong potential for appearance modeling within tracking-by-detection frameworks [13].

Popular real-time MOT frameworks such as DeepSORT [14] and StrongSORT [15] integrate motion prediction with appearance cues to associate object identities across frames. However, their appearance modules typically rely on lightweight CNN-based embeddings trained on limited-scale pedestrian datasets, which predominantly capture low-level visual cues. Consequently, these embeddings often fail under heavy occlusion, drastic illumination changes, viewpoint variations, or scenes containing visually similar individuals, leading to increased identity switch (IDSW) errors. Recent studies [16–18] demonstrate that replacing these shallow, domain-specific embeddings with more expressive models such as CLIP yields substantial improvements in re-identification robustness. CLIP's semantically grounded and context-aware representations enable more reliable identity discrimination in crowded and visually complex scenes, strengthening the appearance modeling backbone of tracking-by-detection pipelines.

Despite the individual successes of YOLOv8 in object detection and CLIP in representation learning, a systematic investigation of their joint integration within state-of-the-art MOT frameworks remains limited. This study aims to fill this gap by evaluating the impact of combining YOLOv8 with CLIP-derived appearance embeddings in DeepSORT and StrongSORT pipelines. Unlike prior works that modify multiple pipeline components simultaneously, this study isolates the effect of appearance feature quality on identity association stability, with a particular focus on reducing IDSW. The evaluation is conducted on two widely used MOT benchmarks, MOT15 [19] and MOT16 [20], which feature dense crowds, frequent occlusions, and complex scene dynamics.

1.1. Research Questions and Motivation

This study investigates the role of appearance representation quality in multi-object tracking systems, particularly within tracking-by-detection pipelines operating in crowded scenes. While prior approaches primarily rely on conventional ReID embeddings for appear-

https://doi.org/10.3390/electronics15020265

ance modeling, recent advances in large-scale vision–language models offer semantically richer feature representations. However, the extent to which such high-capacity embeddings improve identity consistency independent of detector or motion model changes has not been thoroughly quantified. Motivated by this gap, this work isolates the contribution of CLIP-derived appearance features and evaluates their effectiveness in reducing identity switches under challenging tracking conditions. Based on this motivation, we formulate the following research questions:

1. How does replacing conventional ReID appearance embeddings with CLIP-derived representations affect identity consistency in multi-object tracking pipelines?
2. To what extent do semantically rich, high-capacity appearance features contribute to reducing identity switches (IDSW), particularly in crowded and occlusion-heavy scenes?
3. Can improvements in tracking performance be directly attributed to enhanced appearance feature quality when other components of the tracking-by-detection framework are held constant?

1.2. List of Main Contributions

The main contributions of this work are summarized as follows:

1. An enhanced tracking-by-detection architecture is proposed in which CLIP-derived visual embeddings are integrated into the DeepSORT and StrongSORT frameworks, replacing conventional CNN-based ReID models with semantically rich appearance representations.
2. A controlled and systematic analysis of appearance feature quality in multi-object tracking is conducted by isolating the impact of CLIP-based embeddings on identity association stability, with particular emphasis on identity switch (IDSW) reduction in crowded and occluded environments.
3. A high-capacity object detector, YOLOv8x fine-tuned on the MOT20 dataset, is incorporated to examine how improvements in detection accuracy propagate through the tracking pipeline and interact with enhanced appearance modeling.
4. Extensive experimental evaluations are performed on established multi-object tracking benchmarks, including MOT15 and MOT16, assessing tracking performance under challenging conditions characterized by dense crowds, severe occlusions, and complex scene dynamics.
5. The selection of CLIP over alternative vision–language models, such as ALIGN [21], is justified based on the availability of publicly released pretrained weights, large-scale training data, strong cross-domain generalization capability, and reproducibility, establishing its suitability for practical deployment in modern multi-object tracking systems.

The remainder of the paper is organized as follows: Section 2 presents the YOLOv8 detector, CLIP embedding model, and the DeepSORT and StrongSORT algorithms. Section 3 introduces the MOT datasets used. Section 4 defines the evaluation metrics. Section 5 details the experimental setup and discusses results. Section 6 concludes the paper and outlines future research directions.

2. Background

This section provides a comprehensive overview of the key components forming the foundation of the proposed multi-object tracking (MOT) system, including the object detection model (YOLOv8), tracking algorithms (DeepSORT [14] and StrongSORT [15]), and the CLIP model [12]. DeepSORT and StrongSORT are regarded as two of the most influential and foundational association-based trackers in the MOT community.

2.1. Object Detection Models (YOLOv8)

YOLOv8 [7], one of the latest evolutions of the YOLO series, introduces several architectural enhancements. It adopts an anchor-free detection scheme, simplifying the model while improving its generalization to various object sizes. Furthermore, it features a decoupled detection head that separately addresses object classification and localization, contributing to more stable training and improved performance. Additionally, the use of dynamic label assignment during training enables the model to adapt to ambiguous scenarios by adjusting the ground truth assignment strategy. The architecture of YOLOv8 comprises three main components: a backbone for feature extraction, a neck for multi-scale feature aggregation, and a detection head that outputs class probabilities and bounding box coordinates in a single forward pass. This streamlined design facilitates real-time inference while maintaining high detection accuracy. Figure 1 illustrates the overall architecture of YOLOv8, showing the processing flow from input image to final detections via the backbone, neck, and detection head.

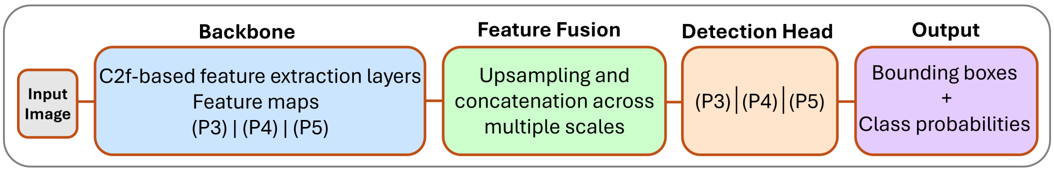


Figure 1. An overview of the YOLOv8 architecture, showing the backbone feature extraction, feature fusion, and multi-scale detection heads (reproduced from [7]).

2.2. Feature Extraction Models for Re-Identification

In MOT, object detection alone is insufficient; the system must also consistently identify and re-identify objects across video frames. This requires strong feature extraction models capable of generating discriminative and robust embeddings for each detected object.

Recent works have increasingly explored vision–language models, particularly CLIP-based representations,to enhance feature learning for tracking and re-identification. CLIP-ReID [22] demonstrates the effectiveness of CLIP's pretrained visual features for appearance embedding in person Re-ID tasks, which are closely related to the appearance models used in tracking association. n parallel, YOLO-World [23] introduces open-vocabulary object detection using vision-language modeling as a detection backbone, enabling category-agnostic detection. Furthermore, several MOT approaches such as VSE-MOT [16], ReTrackVLM [24], and OVTrack [25] incorporate CLIP-based visual features or distillation strategies within transformer-based tracking frameworks.

More recently, zero-shot tracking paradigms have emerged that rely heavily on vision–language models. Z-GMOT [26] proposes zero-shot generic multiple object tracking by leveraging a vision language detector to track unseen object categories without any training, while maintaining a detection pipeline tracker distinct from traditional SORT-based association. Similarly, ReferGPT [27] introduces a zero-shot referring multi-object tracking framework that combines CLIP-based semantic representations with natural-language queries to associate and track multiple targets. Despite their effectiveness, these approaches employ custom association mechanisms rather than classical MOT pipelines.However, none of these approaches explicitly integrate CLIP embeddings into traditional DeepSORT or StrongSORT pipelines for appearance modeling and data association, which remains an unexplored direction that our work addresses. Table 1 summarizes key features of these vision–language MOT methods, including CLIP usage, detection, tracking paradigms, purpose, and alternative tracking approaches.

Table 1. Comparison of Vision–Language Based Methods for Multi-Object Tracking.

Year	Method/Paper	Uses CLIP	Uses Object Detection	Tracking Paradigm	Main Purpose	Tracking Method Used (Instead of DeepSORT/StrongSORT)
2023	OVTrack: Open-Vocabulary Multiple Object Tracking	Yes (CLIP feature distillation)	Yes	Detection + tracking head	Track seen and unseen categories	Open-vocabulary MOT framework: learned tracking head (no Kalman filter or Hungarian matching)
2023	CLIP-ReID: Exploiting Vision-Language Model for Image Re-Identification	Yes (CLIP visual features)	No	Feature learning (Re-ID only)	Improve appearance embeddings	No tracking: Re-ID model only (can be plugged into DeepSORT/StrongSORT)
2024	YOLO-World: Real-Time Open-Vocabulary Object Detection	Yes (vision-language detector)	No	Object detection	Open-vocabulary detection	No tracking: detector only (can feed DeepSORT/StrongSORT)
2024	Z-GMOT: Zero-shot Generic Multiple Object Tracking	Yes (leverages a vision-language detector iGLIP/GLIP)	Yes	Tracking-by-Detection (zero-shot generic MOT)	Track multiple generic unseen objects without predefined categories	MA-SORT—motion & appearance association specifically designed for generic object association (replaces typical SORT/DeepSORT)
2025	ReTrackVLM: Transformer-Enhanced MOT with Cross-Modal Embeddings and Zero-Shot Re-ID Integration	Yes (VLM/CLIP-style embeddings)	Yes	Transformer-based MOT	Enhance identity association	Transformer encoder–decoder tracking: end-to-end learned association
2025	VSE-MOT: Multi-Object Tracking in Low-Quality Video Scenes Guided by Visual Semantic Enhancement	Yes (CLIP image encoder)	Yes	Tracking-by-detection	Improve robustness in low-quality videos	Transformer-based MOT, query-based tracking with learned association
2025	ReferGPT: Towards Zero-Shot Referring Multi-Object Tracking	Yes—uses CLIP-based semantic encoding for matching generated captions with queries	Yes	Tracking-by-Detection (with Kalman filter association)	Track objects specified by natural language queries in a zero-shot manner	Kalman filter + fuzzy query matching within a tracking-by-detection pipeline (not DeepSORT/StrongSORT)

In this study, we explore the use of the CLIP model as a novel feature extractor. CLIP, developed by OpenAI [12], for its widespread applications in many vision problems [28–30], learns aligned image-text representations using contrastive learning on a large-scale dataset. Unlike traditional models trained on labeled images alone, CLIP leverages natural language supervision to learn semantically rich and generalizable visual embeddings. This characteristic makes CLIP particularly suitable for re-identification tasks, where subtle visual differences between similar-looking objects are important. CLIP employs two separate encoders: a visual encoder for processing images and a text encoder for language inputs. These encoders are jointly trained to project inputs into a shared latent space where paired image text embeddings are close in cosine similarity. For MOT applications, only the visual encoder is used to produce object embeddings for tracking. Figure 2 illustrates the CLIP model architecture, showing the dual encoder design and contrastive learning objective.

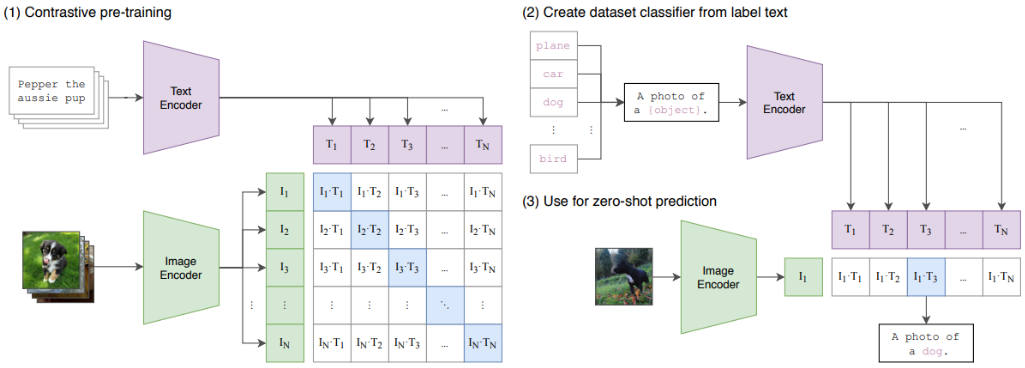

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Figure 2. Overview of CLIP approach: CLIP is different from traditional image models because it learns from both images and text at the same time. It trains an image encoder and a text encoder to match images with their related text descriptions. Instead of using a fixed classifier, CLIP uses the text encoder during testing to create a zero-shot classifier by embedding the names or descriptions of the target classes. [12].

In addition, we consider baseline feature extractors commonly used in MOT literature. DeepSORT employs the MARS CNN [31], originally designed for person re-identification, as its default appearance model. StrongSORT integrates the OSNet [32], which introduces omni-scale feature learning through specialized residual blocks capable of capturing multi-scale representations. These baselines provide meaningful comparisons against the CLIP-based embedding pipeline in terms of both accuracy and computational efficiency.

2.3. Tracking Models: DeepSORT and StrongSORT

Following the introduction of object detection and feature extraction modules, this section presents the multi-object tracking frameworks employed in this study, namely, DeepSORT [14] and StrongSORT [15].

2.3.1. DeepSORT

DeepSORT [14] is an enhanced tracking by detection framework that extends the original SORT algorithm [14]. It incorporates deep appearance feature embeddings to establish robust associations between object detections across frames, enabling accurate tracking even under challenging conditions such as occlusion, re-identification, and partial visibility. The architecture of DeepSORT comprises three primary components, as illustrated in Figure 3. First, a deep appearance descriptor is extracted for each detected object, providing a robust representation that facilitates re-identification across frames. These descriptors are crucial for distinguishing between visually similar objects and maintaining consistent identities over time. Second, a Kalman filter is used to estimate the state (position and velocity) of each tracked object, predicting its location in subsequent frames. This is particularly effective when detections are noisy or intermittent due to occlusion. Third, the Hungarian algorithm is employed for data association, solving the assignment problem between new detections and existing tracks based on a combination of motion (from the Kalman filter) and appearance similarity. If a detection cannot be matched to any existing track, a new track is initialized. Conversely, tracks that remain unmatched for a predefined number of frames are terminated. DeepSORT's balance of efficiency and accuracy has led to its widespread use in applications such as surveillance, autonomous vehicles, and crowd analytics.

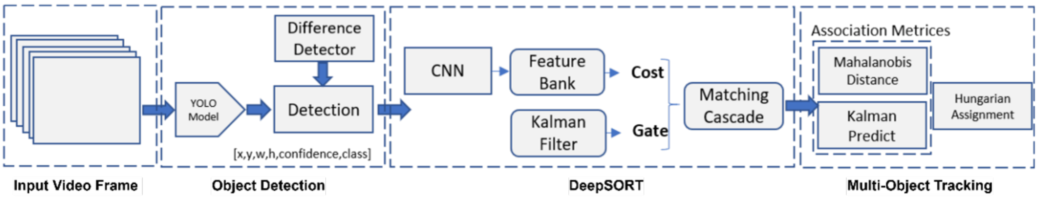


Figure 3. DeepSORT architecture [8].

2.3.2. StrongSORT

StrongSORT [15] is an efficient algorithm for multi-object tracking, designed to overcome the limitations observed in earlier tracking methods. A key distinguishing feature of StrongSORT is its integration of deep neural networks for extracting appearance-robust features that remain consistent under significant visual variations. As illustrated in Figure 4, StrongSORT extends the DeepSORT framework by incorporating additional modules for more reliable tracking. It introduces two core innovations: the Appearance Free Link (AFLink) module and the Gaussian Smoothed Interpolation (GSI) technique. AFLink facilitates matching between fragmented object trajectories without relying entirely on appearance features, thereby reducing computational dependency and improving robustness against occlusions. Meanwhile, GSI estimates the position of missing detections through interpolation, enhancing continuity in object trajectories. StrongSORT also leverages high-performance object detectors and advanced feature extractors (such as OSNet) to maintain trajectory consistency across frames.

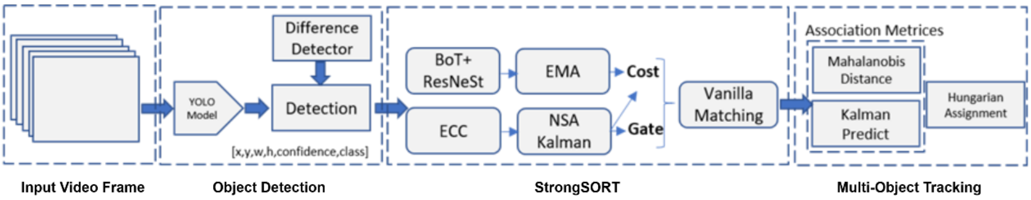


Figure 4. StrongSORT architecture [8].

3. Benchmark Datasets

To evaluate the effectiveness of our proposed tracking methods, we utilize three widely adopted multi-object tracking benchmark datasets: MOT15 [19], MOT16 [20], and MOT20 [33]. These datasets provide varying degrees of complexity in terms of crowd density, occlusion, lighting, and camera motion. Figure 5 illustrates sample frames from the three datasets: Figure 5a shows scenes from MOT15, Figure 5b presents MOT16 scenes, and Figure 5c displays MOT20 scenes [19,20,33].

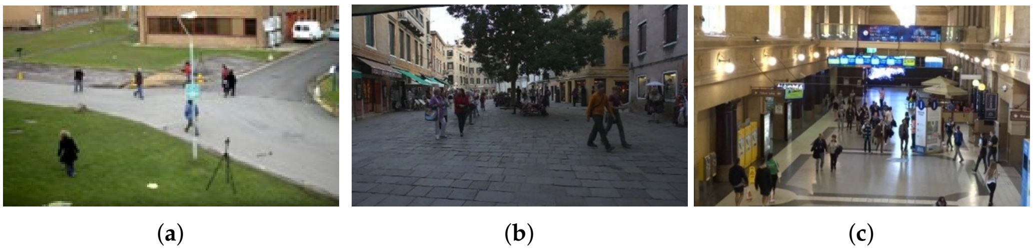


Figure 5. Example frames from the MOT15, MOT16, and MOT20 datasets [19,20,33]. (**a**) MOT15 sample scenes. (**b**) MOT16 sample scenes. (**c**) MOT20 sample scenes.

3.1. MOT15 Dataset

The MOT15 [19] dataset contains 22 video sequences recorded in both indoor and outdoor environments, totaling 11,297 annotated detections. This dataset presents several key challenges, including frequent occlusions, varying camera viewpoints, and complex crowd motion, making it suitable for evaluating the robustness of tracking algorithms. MOT15 includes bounding box annotations, object IDs, and visibility ratios, enabling comprehensive performance evaluation under various difficulty levels.

3.2. MOT16 Dataset

MOT16 [20] is an enhanced and standardized extension of MOT15. It consists of 14 video sequences and over 1100 annotated object trajectories. The dataset adheres to consistent annotation guidelines and spans a wide range of environments, such as urban streets, pedestrian zones, and shopping malls. MOT16 introduces more complex challenges like dense occlusions, motion camouflage, and variable lighting. Performance evaluation is conducted using metrics including Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and the Identity F1 Score (IDF1), providing a comprehensive assessment of accuracy, precision, and identity consistency.

3.3. MOT20 Dataset

The MOT20 [33] dataset introduces a higher level of complexity, featuring 8 ultra-dense video sequences with 1564 unique identities and over 4 million bounding box annotations. This dataset focuses on extremely crowded scenes with severe occlusions and intricate interactions among pedestrians. It includes detailed annotations for bounding boxes, object IDs, occlusion levels, and visibility, offering a rigorous benchmark for evaluating the robustness of tracking systems in real-world congested environments.

4. Evaluation Metrics

In the domain of MOT, several evaluation metrics are widely used to assess different aspects of a tracking system's performance. In this study, we utilize five core metrics: Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Identity Switches (IDSW), Precision, and Recall. Each of these metrics is detailed below.

4.1. Multiple Object Tracking Accuracy (MOTA)

MOTA evaluates the overall tracking accuracy by considering false positives, false negatives, and identity switches during tracking [34–36]. A higher MOTA value indicates better performance in minimizing detection and identification errors.

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

where FN represents the false negatives, FP represents the false positives, IDSW represents the identity switches, and GT represents the ground truth objects.

4.2. Multiple Object Tracking Precision (MOTP)

MOTP measures the accuracy of object localization by evaluating how well the predicted positions align with the ground truth locations [34–36]. A higher MOTP score reflects better localization precision.

$$\text{MOTP} = \frac{1}{|TP|} \sum_{i=1}^{|TP|} S_i \tag{2}$$

where S_i is the similarity score (typically IoU) for each true-positive match.

4.3. Identity Switches (IDSW)

The IDSW metric quantifies the number of times a tracked object's identity is incorrectly reassigned during the tracking process [15,35–37,37]. Fewer identity switches indicate a better ability to maintain object identity continuity, which is critical for applications like surveillance or autonomous systems.

$$\text{IDSW} = \sum_{t=1}^{T} IDSW_t \tag{3}$$

where $IDSW_t$ is the number of identity switches that occur at frame t. A switch happens when the tracker changes the ID assigned to the same ground-truth object from one frame to the next.

4.4. Precision

Precision evaluates the proportion of correctly predicted positive instances among all positive predictions made by the model. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

where TP is the number of true positives and FP is the number of false positives.

4.5. Recall

Recall measures the proportion of correctly predicted positive instances among all actual positive instances in the dataset:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

where TP is the number of true positives and FN is the number of false negatives.

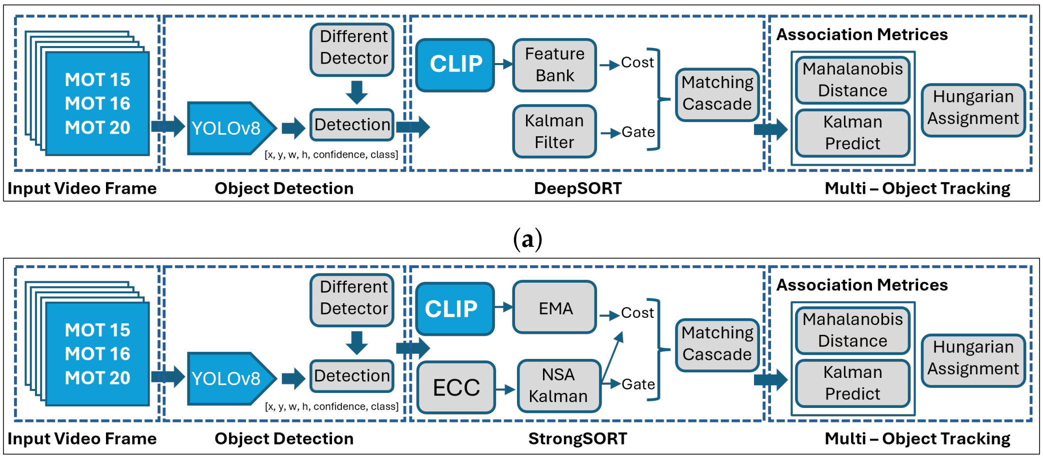
5. Experiments and Results

5.1. Proposed Architectures

This work proposes an enhanced tracking-by-detection architecture based on the DeepSORT and StrongSORT frameworks, in which both the object detector and the appearance embedding modules are explicitly upgraded. The primary objective of the proposed architecture is to improve identity association robustness by strengthening appearance feature quality while preserving the original motion and association mechanisms of the baseline trackers. Specifically, the CLIP vision encoder is integrated into the appearance extraction stage of both trackers to replace the default CNN-based ReID models, and the YOLOv8 detector is adopted in place of the original detection backbones.

Figure 6 illustrates the overall processing pipeline. Given an input video sequence, each frame is first passed to the YOLOv8 detector, which outputs bounding boxes, confidence scores, and class labels for detected objects. These detections are then forwarded to the tracking stage for identity association across frames. For each detected bounding box, the corresponding image crop is extracted and resized to match the input resolution required by the CLIP vision encoder. The CLIP model is used in inference mode only, and no fine-tuning is performed during tracking. The resulting high-dimensional embeddings are L2-normalized before being used in the data association process.

In the DeepSORT-based architecture, the CLIP-derived appearance embeddings are stored in a feature gallery (feature bank) associated with each active track. During data association, cosine distance is computed between the current detections and existing track features, and this appearance cost is combined with motion-based gating derived from the Kalman filter. This design preserves the original DeepSORT matching strategy while explicitly isolating the impact of enhanced appearance features on identity association stability.



(**a**)

(**b**)

Figure 6. The integration of the CLIP model into the two trackers: DeepSORT (**a**) and StrongSORT (**b**). (**a**) CLIP model integrated in the DeepSORT pipeline. (**b**) CLIP model integrated in the StrongSORT pipeline.

In the StrongSORT-based architecture, the CLIP embeddings are incorporated into the exponential moving average (EMA) module, which maintains a temporally smoothed representation of each track's appearance. These EMA-updated features are used to compute the appearance affinity matrix, which is subsequently combined with motion and geometric constraints during the Hungarian matching stage. By integrating CLIP features into the EMA mechanism, StrongSORT benefits from both semantically rich embeddings and temporal appearance smoothing.

Importantly, aside from the detector and appearance embedding modules, all other components of the DeepSORT and StrongSORT pipelines—including Kalman filtering, motion prediction, gating thresholds, and matching logic—are kept identical to their official implementations. This design choice ensures that any observed performance differences can be attributed directly to improvements in appearance representation and detection quality.

Three publicly available MOT benchmarks were used for evaluation: MOT15, MOT16, and MOT20. The YOLOv8 detector was employed under two experimental settings. In the first setting, YOLOv8 with official pretrained weights was directly integrated into both tracking pipelines and evaluated on MOT15 and MOT16. In the second setting, YOLOv8 was fine-tuned on MOT20 to enhance detection performance in crowded scenes and subsequently cross-evaluated on MOT15 and MOT16 to assess generalization. Across all experiments, the effect of replacing conventional ReID embeddings with CLIP-derived features was systematically analyzed, with a particular focus on identity switch reduction and tracking stability. The quantitative results and ablation analyses are presented in the subsequent sections.

5.2. Experimental Setups

All experiments were conducted in Python 3.11 and executed in Google Colab, utilizing a single NVIDIA RTX A100 GPU. The object detection module employed the YOLOv8 architecture [7], which was selected for its proven accuracy and continued open-source availability. Specifically, YOLOv8x (non-tuned) and YOLOv8x (tuned) were chosen. For non-tuned YOLOv8x, weights from the COCO dataset [37] were used, while for the tuned version, YOLOv8x was trained on the MOT20 [33] benchmark for 100 epochs, learning rate 0.01, image size 640 × 640. The obtained weights were evaluated on the MOT15 [19] and MOT16 [20] benchmarks to test the generalization capabilities, considering the crowded and challenging scenes present in the MOT20 benchmark. For tracking, both DeepSORT [14] and StrongSORT [15] were utilized with settings similar to their main repositories. The baseline experiment contains the official codes that were utilized directly. The proposed experiment integrated the official CLIP model [12] as the feature extractor model to replace the default one in the DeepSORT and StrongSORT models. For consistency and reproducibility, default hyperparameters were retained for both DeepSORT and StrongSORT across all experiments.

5.3. Quantitative Results

In this section, the quantitative results are presented in a clear and systematic manner to highlight the impact of the proposed improvements. Specifically, three sets of experiments were conducted to evaluate the contributions of the CLIP model and YOLOv8x detector in enhancing identity consistency and tracking performance for DeepSORT and StrongSORT. Each experiment is designed to isolate the effect of feature embeddings, detector replacement, and generalization across datasets.

The first experiment evaluates the effect of integrating CLIP embeddings into both trackers to enhance identity preservation. The second experiment examines the impact of replacing the default detector with the official YOLOv8x model, assessing improvements in detection quality. The third experiment investigates the generalization capability of YOLOv8x, trained on the MOT20 benchmark and tested on MOT15 and MOT16, to evaluate cross-dataset performance.

5.3.1. DeepSORT-Experiment: Impact of CLIP

To clearly demonstrate the generalization of these findings, DeepSORT [14] was evaluated on the MOT15 benchmark. Table 2 summarizes the results for three key scenarios: (1) adding CLIP embeddings, (2) replacing the detector with the official YOLOv8x, and (3) using fine-tuned YOLOv8x weights from MOT20.

Key observations from the results include:

1. On PETS09-S2L1, integrating CLIP reduced ID switches from 19 to 17, demonstrating improved identity tracking.
2. On ETH-Sunnyday, using fine-tuned YOLOv8x weights improved both MOTP and Precision, indicating better localization and detection quality.

These results clearly demonstrate that integrating CLIP embeddings enhances identity consistency, while high-capacity detectors such as YOLOv8x further boost overall tracking performance. The combination of robust embeddings and accurate detection consistently improves tracking metrics across sequences.

Table 2. DeepSORT results on the MOT15 benchmark. With and without CLIP on sequence PETS09-S2L1, addition of official YOLO on sequence ETH-Sunnyday, and ★ represents using the finetuned weights of YOLOv8 on the MOT20 benchmark.

Seq. ID	Exp.	Recall ↑	Prec. ↑	IDSW ↓	MOTA ↑	MOTP ↑
PETS09-S2L1	**NO CLIP**	94.00	92.50	19.00	86.00	23.30
	With CLIP	93.70	92.30	**17.00**	85.60	**23.80**
ETH-Sunnyday	**No CLIP**	94.10	74.30	2.00	61.40	17.40
	With CLIP	93.90	74.30	2.00	61.30	**17.80**
ETH-Sunnyday ★	**No CLIP**	93.20	86.80	1.00	79.00	15.20
	With CLIP	93.10	**86.90**	1.00	79.00	**15.60**

5.3.2. StrongSORT-Experiment 1: Impact of CLIP

In the first experiment, StrongSORT [15] was benchmarked on the MOT16 [20] dataset to systematically evaluate the impact of integrating the CLIP model into the appearance feature extractor. Table 3 presents a clear comparison of the evaluation metrics Recall, Precision, IDSW, MOTA, and MOTP between the original StrongSORT and the CLIP-enhanced version across three sequences.

The results demonstrate consistent improvements across one or more metrics. Notably, in Sequence 02, the IDSW decreased from 479 to 318, corresponding to a reduction of approximately 33.6%, while Precision, Recall, MOTA, and MOTP also showed measurable gains. For Sequences 09 and 11, reductions in IDSW and increases in MOTA indicate that integrating CLIP significantly improves identity stability in crowded or challenging frames.

These results clearly highlight that the CLIP-enhanced StrongSORT model achieves more stable and accurate tracking performance, particularly in sequences with high identity ambiguity, demonstrating the effectiveness of CLIP embeddings in maintaining identity consistency.

Table 3. StrongSORT: With and without CLIP. Results on the MOT16 benchmark.

Seq. ID	Exp.	Recall ↑	Prec. ↑	IDSW ↓	MOTA ↑	MOTP ↑
02	**No CLIP**	97.30	85.30	479.00	77.90	91.30
	With CLIP	**97.50**	**85.60**	**318.00**	**79.30**	**91.80**
09	**No CLIP**	92.30	80.10	45.00	68.50	96.40
	With CLIP	**93.50**	80.00	**42.00**	**69.30**	96.30
11	**No CLIP**	91.70	92.20	50.00	83.30	96.40
	With CLIP	**92.00**	92.00	**49.00**	**83.50**	96.30

5.3.3. StrongSORT-Experiment 2: Cross-Dataset Generalization of YOLO

The second experiment evaluates the effect of replacing the default StrongSORT detector (YOLOX [38]) with YOLOv8x, and studies its generalization when trained on MOT20 [33] but evaluated on MOT15 [19]. Table 4 presents three key scenarios: (1) the impact of CLIP on MOT15 sequence PETS09-S2L1, (2) the effect of YOLOv8x integration on sequence TUD-Stadmittee, and (3) cross-dataset evaluation of fine-tuned YOLOv8x weights.

The main findings from these experiments are summarized as follows:

1. CLIP integration on PETS09-S2L1 reduced IDSW by approximately 60%, demonstrating significant improvements in identity consistency.

2. Replacing the default detector with YOLOv8x on TUD-Stadmittee reduced IDSW by 44%, confirming the benefits of high-capacity detection for tracking performance.
3. Fine-tuned YOLOv8x generalized well to MOT15, improving MOTA (78.70 → 82.79) and MOTP (22.40 → 26.70). Adding CLIP further reduced IDSW from 15 to 10 (33% improvement), illustrating the complementary effect of robust embeddings combined with accurate detection.

Overall, these results clearly demonstrate that integrating high-capacity detection with CLIP embeddings significantly enhances both identity preservation and overall tracking performance across sequences and generalizes effectively to cross-dataset evaluation.

Table 4. StrongSORT with YOLOv8: With and without CLIP. Results on the MOT15 benchmark. ★ represents using the finetuned weights of YOLOv8 on the MOT20 benchmark.

Seq. ID	Exp.	Recall ↑	Prec. ↑	IDSW ↓	MOTA ↑	MOTP ↑
PETS09-S2L1	**NO CLIP**	94.60	89.30	100.00	81.10	24.30
	With CLIP	92.70	**92.20**	**42.00**	**84.00**	24.20
TUD-Stadmittee	**No CLIP**	82.40	92.60	9.00	78.70	22.40
	With CLIP	**82.80**	96.50	**5.00**	**79.30**	**22.90**
TUD-Stadmittee ★	**No CLIP**	86.20	97.60	15.00	82.70	26.70
	With CLIP	88.20	96.20	**10.00**	**83.90**	**26.80**

5.4. Qualitative Results

This section presents qualitative results that visually support and clarify the quantitative improvements discussed earlier. In particular, the figures provide an intuitive comparison of tracking behavior with and without CLIP integration, allowing the reader to directly observe reductions in identity switches (IDSW) across challenging scenarios.

Figure 7 illustrates representative frames from different MOT16 sequences under crowded and dynamic conditions, highlighting how CLIP-enhanced appearance features lead to more consistent identity assignment over time. These visual examples are intended to complement numerical metrics by demonstrating practical tracking improvements in real-world scenes.

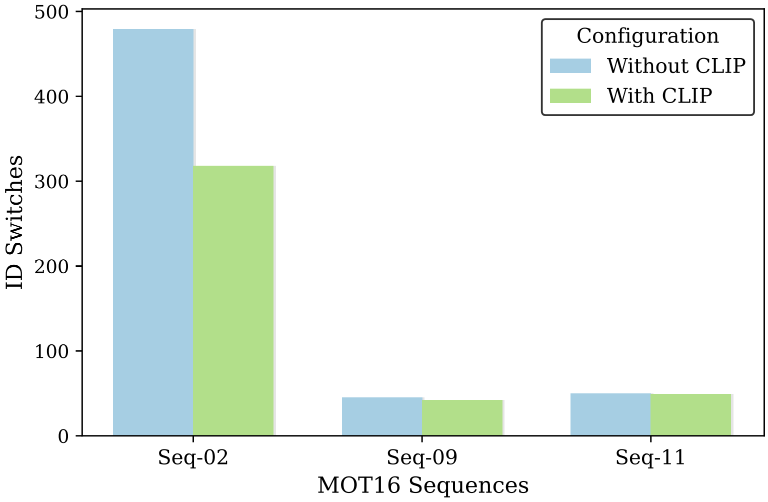


Figure 7. StrongSORT without integrating YOLOv8: Impact of CLIP on IDSW on MOT16 Benchmark.

https://doi.org/10.3390/electronics15020265

5.4.1. Graphical Analysis of Results

Figure 7 (shown above) compares StrongSORT performance on MOT16 sequences with and without CLIP. The bar chart clearly shows a substantial reduction in ID switches when CLIP is enabled, most notably in Sequence 02, which contains heavy occlusions and dense interactions. More moderate but consistent improvements are also observed in Sequences 09 and 11, reinforcing the robustness of CLIP integration across varying levels of scene complexity.

Figure 8 provides a clear visual comparison of DeepSORT and StrongSORT on the MOT15 PETS09-S2L1 sequence, illustrating the impact of CLIP integration on identity switches (IDSW). The bar chart contrasts each tracker's performance with and without CLIP, enabling a direct and intuitive assessment of identity consistency.

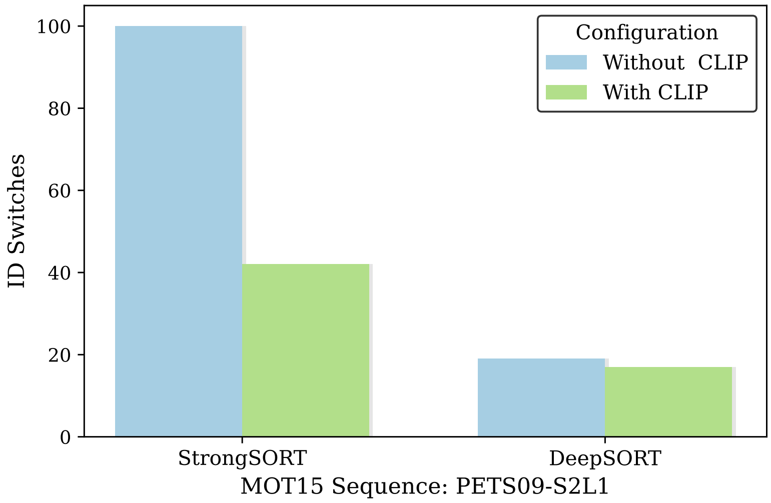


Figure 8. StrongSORT & DeepSORT with the integration of YOLOv8: Impact of CLIP on IDSW on MOT15 Benchmark.

The results demonstrate that integrating CLIP reduces ID switches for both trackers, confirming its effectiveness in enhancing appearance-based association. Notably, the improvement is substantially more pronounced for StrongSORT, where IDSW is reduced by more than half, indicating that CLIP features strongly complement StrongSORT's association mechanism. In contrast, DeepSORT exhibits a smaller but consistent reduction, reflecting its more limited reliance on appearance embeddings.

Overall, this figure clearly highlights how CLIP integration improves tracking robustness, particularly for StrongSORT, and reinforces the conclusion that richer visual representations are especially beneficial in crowded and visually ambiguous tracking scenarios.

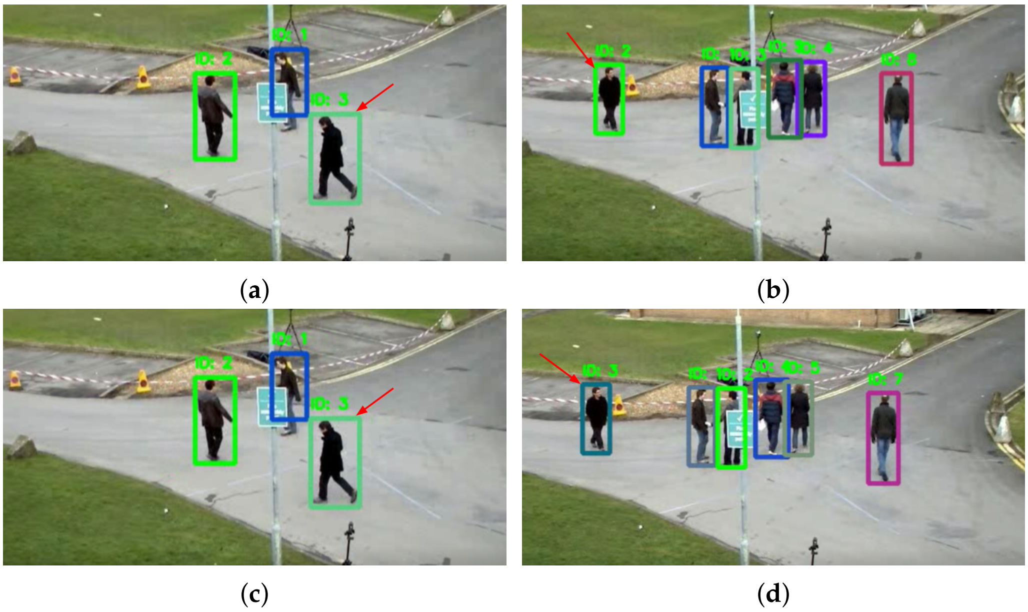
5.4.2. Visual Results on MOT15 and MOT16

This section presents qualitative visual results on the MOT15 and MOT16 benchmarks to clearly illustrate the impact of integrating the CLIP model into the StrongSORT tracker. Specifically, the PETS09-S2L1 sequence from MOT15 and Sequence 02 from MOT16 are selected due to their dense interactions and frequent occlusions, making them representative and challenging test cases for identity association. The effect of CLIP integration is analyzed by visually comparing identity switch (IDSW) behavior before and after its introduction.

Figures 9 and 10 provide frame-by-frame visual comparisons that allow direct observation of tracking behavior over time. When CLIP is integrated, object identities are preserved consistently across frames, even under occlusion and close interactions. In contrast, the configurations without CLIP exhibit frequent identity fragmentation and multiple ID switches, which are highlighted in the visual examples. These qualitative results clearly demonstrate

https://doi.org/10.3390/electronics15020265

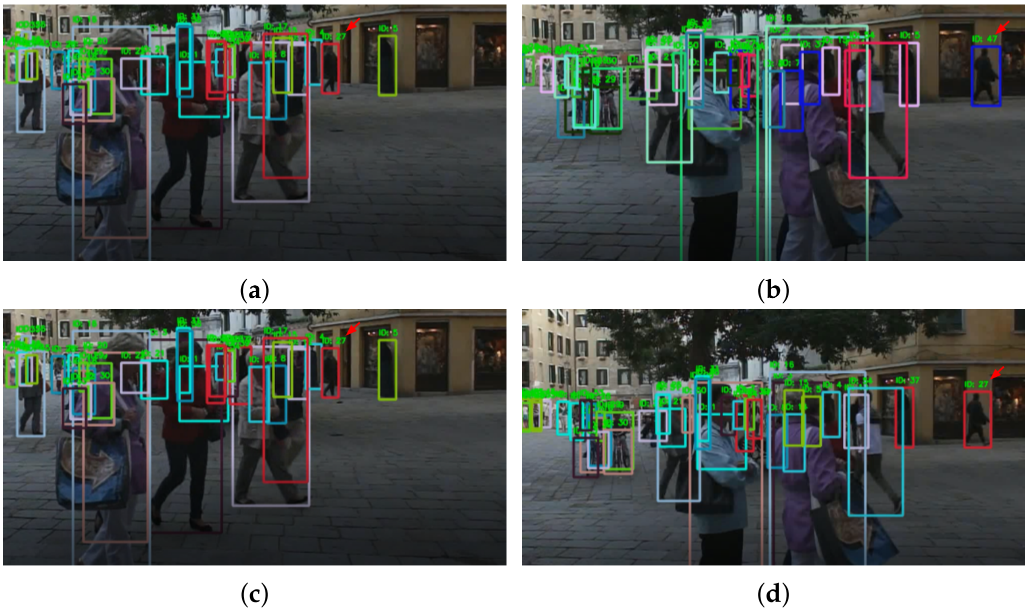
how CLIP-based embeddings improve appearance discrimination and temporal identity consistency in crowded environments.



(**a**) (**b**)

(**c**) (**d**)

Figure 9. Impact of integrating the CLIP model into the StrongSORT tracker in terms of IDSW on the MOT15 benchmark sequence PETS09-S2L1. The red arrow indicates the ID of a tracked person, which switches from 3 to 2 without the CLIP model, but remains consistent when the CLIP model is applied. (**a**) No-CLIP before ID 3 switches. (**b**) No-CLIP after ID 3 switches. (**c**) With-CLIP before: ID 3. (**d**) With-CLIP after: ID 3.



(**a**) (**b**)

(**c**) (**d**)

Figure 10. Impact of integrating the CLIP model into the StrongSORT tracker in terms of IDSW on the MOT16 benchmark sequence 02. The red arrow indicates the ID of a tracked person, which switches from 27 to 47 without the CLIP model, but remains consistent when the CLIP model is applied. (**a**) No-CLIP before ID 27 switches. (**b**) No-CLIP after ID 27 switches. (**c**) With-CLIP before: ID 27. (**d**) With-CLIP after: ID 27.

In summary, both the qualitative visual evidence and the quantitative metrics consistently confirm that CLIP integration significantly enhances identity association, substantially reduces ID switches, and effectively complements the high-capacity detection provided by YOLOv8x. Together, these improvements lead to more robust and reliable multi-object tracking performance across diverse and challenging scenarios.

6. Conclusions

This work investigated a systematic integration of CLIP-based appearance embeddings and a fine-tuned YOLOv8x detector into the DeepSORT and StrongSORT frameworks, introducing two methodological contributions to multi-object tracking. First, we replaced traditional CNN-based ReID descriptors with CLIP's semantically enriched visual embeddings, enabling a substantially more discriminative appearance model that directly targets one of the fundamental weaknesses of real-time trackers—identity switches in crowded scenes. Second, we incorporated a high-capacity detector (YOLOv8x), fine-tuned on MOT20 to improve detection quality while evaluating its cross-dataset generalization on MOT15 and MOT16. Together, these innovations provide a principled analysis of how upgrading both the detection and appearance modules impacts the stability and robustness of MOT systems.

Experimental results confirm that CLIP embeddings significantly reduce IDSW across benchmarks, and that YOLOv8x retains strong performance even when transferred across datasets with varying crowd densities. These findings establish a clear link between enhanced feature quality, stronger detections, and improved tracking continuity. Future extensions will focus on fine-tuning CLIP for MOT-specific domains and incorporating temporal/contextual cues to further address remaining identity inconsistencies in highly congested environments.

Author Contributions: Conceptualization, K.A., A.S.Y. and H.M.; methodology, K.A. and A.S.Y.; software, K.A. and A.S.Y.; formal analysis, K.A.; investigation, K.A. and A.S.Y.; resources, K.A. and A.S.Y.; data curation, K.A.; writing—original draft preparation, K.A. and A.S.Y.; writing—review and editing, K.A., A.S.Y. and H.M.; visualization, K.A.; supervision, H.M.; project administration, K.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in https://motchallenge.net (accessed on 1 September 2025), and reference [14].

Acknowledgments: Khadijah Alkandary and Ahmet Serhat Yildiz express their heartfelt gratitude to their family for their endless support, patience, and encouragement throughout this research journey. Also, Ahmet Serhat Yildiz's Ph.D. is sponsored by the Ministry of National Education of Türkiye.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
2. Yeung, S.; Downing, N.L.; Fei-Fei, L.; Milstein, A. Bedside computer vision—moving artificial intelligence from driver assistance to patient safety. *N. Engl. J. Med.* **2018**, *378*, 1271–1273. [CrossRef] [PubMed]
3. Li, J.; Wang, B.; Ma, H.; Gao, L.; Fu, H. Visual feature extraction and tracking method based on corner flow detection. *ICCK Trans. Intell. Syst.* **2024**, *1*, 3–9. [CrossRef]
4. Ravindran, R.; Santora, M.; Jamali, M. Multi object detection and tracking, based on DNN, for autonomous vehicles: A review. *IEEE Sens. J.* **2020**, *21*, 5668–5677. [CrossRef]
5. Coifman, B.; Beymer, D.; McLauchlan, P.; Malik, J. A real-time computer vision system for vehicle tracking and traffic surveillance. *Transp. Res. Part C Emerg. Technol.* **1998**, *6*, 271–288. [CrossRef]
6. Sophokleous, A.; Christodoulou, P.; Doitsidis, L.; Chatzichristofis, S.A. Computer vision meets educational robotics. *Electronics* **2021**, *10*, 730. [CrossRef]
7. Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-time flying object detection with YOLOv8. *arXiv* **2023**, arXiv:2305.09972. [CrossRef]

https://doi.org/10.3390/electronics15020265

8. Alkandary, K.; Yildiz, A.S.; Meng, H. *A Comparative Study of YOLO Series (v3–v10) with DeepSORT and StrongSORT: A Real-Time Tracking Performance Study*; Technical Report; Department of Electronic and Electrical Engineering, Brunel University: London, UK, 2025.
9. Danilowicz, M.; Kryjak, T. Real-Time Multi-object Tracking Using YOLOv8 and SORT on a SoC FPGA. In Proceedings of the International Symposium on Applied Reconfigurable Computing, Seville, Spain, 9–11 April 2025; Springer: Berlin/Heidelberg, Germany, 2025; pp. 214–230.
10. Yu, X.; Liu, X.; Liang, G. YOLOv8-SMOT: An Efficient and Robust Framework for Real-Time Small Object Tracking via Slice-Assisted Training and Adaptive Association. *arXiv* **2025**, arXiv:2507.12087.
11. Liu, Y.; Shen, S. Vehicle Detection and Tracking Based on Improved YOLOv8. *IEEE Access* **2025**, *13*, 24793–24803. [CrossRef]
12. Radford, A.; Kim, J.W.; Hallacy, J.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
13. Zhu, H.; Lu, Q.; Xue, L.; Zhang, P.; Yuan, G. Vision-language tracking with CLIP and interactive prompt learning. *IEEE Trans. Intell. Transp. Syst.* **2024**, *26*, 3659–3670. [CrossRef]
14. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: New York, NY, USA, 2017; pp. 3645–3649. [CrossRef]
15. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. StrongSORT: Make DeepSORT Great Again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [CrossRef]
16. Du, J.; Xing, W.; Li, M.; Yu, F.R. VSE-MOT: Multi-Object Tracking in Low-Quality Video Scenes Guided by Visual Semantic Enhancement. *arXiv* **2025**, arXiv:2509.14060.
17. Asperti, A.; Naldi, L.; Fiorilla, S. An Investigation of the Domain Gap in CLIP-Based Person Re-Identification. *Sensors* **2025**, *25*, 363. [CrossRef] [PubMed]
18. Yang, X.; Gao, X.; Niu, S.; Zhu, F.; Feng, G.; Qu, X.; Camacho, D. CLIP4VI-ReID: Learning Modality-shared Representations via CLIP Semantic Bridge for Visible-Infrared Person Re-identification. *arXiv* **2025**, arXiv:2511.10309.
19. Leal-Taix'e, L.; Milan, A.; Reid, I.; Roth, S.; Schindler, K. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv* **2015**, arXiv:1504.01942. [CrossRef]
20. Milan, A.; Leal-Taix'e, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi object tracking. *arXiv* **2016**, arXiv:1603.00831. [CrossRef]
21. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 4904–4916.
22. Li, S.; Sun, L.; Li, Q. Clip-ReID: Exploiting Vision–Language Model for Image Re-Identification Without Concrete Text Labels. *Proc. Aaai Conf. Artif. Intell.* **2023**, *37*, 1405–1413. [CrossRef]
23. Cheng, T.; Luo, X.; Zhao, H.; Zhou, Y.; Yan, S. YOLO-World: Real-Time Open-Vocabulary Object Detection. *arXiv* **2024**, arXiv:2401.17270.
24. Bayraktar, E. ReTrackVLM: Transformer-Enhanced Multi-Object Tracking with Cross-Modal Embeddings and Zero-Shot Re-Identification Integration. *Appl. Sci.* **2025**, *15*, 1907. [CrossRef]
25. Wu, Y.; Li, Y.; Sheng, H.; Zhang, Z. OVTrack: Open-Vocabulary Multiple Object Tracking. arXiv **2023**, arXiv:2304.08963. [CrossRef]
26. Huang, L.; Wu, Y.; Li, Y.; Sheng, H.; Zhang, Z. Z-GMOT: Zero-Shot Generic Multiple Object Tracking. *arXiv* **2023**, arXiv:2305.17648.
27. Chen, Y.; Sheng, H.; Li, Y.; Zhang, J.; Zhang, Z. ReferGPT: Towards Zero-Shot Referring Multi-Object Tracking. *arXiv* **2025**, arXiv:2504.09195.
28. Li, H.; Zhao, F.; Xue, F.; Wang, J.; Liu, Y.; Chen, Y.; Wu, Q.; Tao, J.; Zhang, G.; Xi, D.; et al. Succulent-YOLO: Smart UAV-Assisted Succulent Farmland Monitoring with CLIP-Based YOLOv10 and Mamba Computer Vision. *Remote Sens.* **2025**, *17*, 2219. [CrossRef]
29. Lin, J.; Gong, S. Gridclip: One-stage object detection by grid-level clip representation learning. *arXiv* **2023**, arXiv:2303.09252. [CrossRef]
30. Vidit, V.; Engilberge, M.; Salzmann, M. Clip the gap: A single domain generalization approach for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 3219–3229.
31. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. MARS: A video benchmark for large-scale person re-identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 868–884.
32. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3702–3712.

33. Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; Leal-Taix'e, L. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv* **2020**, arXiv:2003.09003. [CrossRef]
34. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
35. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 17–35.
36. Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.H.; Geiger, A.; Leal-Taix'e, L.; Leibe, B. HOTA: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 548–578. [CrossRef] [PubMed]
37. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Doll'ar, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
38. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/electronics15020265