Depth and Segmentation Aware frameworks for Multiple Object Tracking

Milad Khanchi

A Thesis

In The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Applied Science (Electrical and Computer Engineering) at

Concordia University

Montréal, Québec, Canada

June 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Milad Khanchi**

Entitled: **Depth and Segmentation Aware frameworks for Multiple Object Tracking**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Krzysztof Skonieczny

_____ External Examiner
Dr. Ali Ayub

_____ Examiner
Dr. Krzysztof Skonieczny

_____ Supervisor
Dr. Maria Amer

_____ Co-supervisor
Dr. Charalambos Poullis

Approved by _____
Dr. Jun Cai, Graduate Program Director

_____ 2025 _____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Depth and Segmentation Aware frameworks for Multiple Object Tracking

Milad Khanchi

Multi-Object Tracking (MOT) remains a challenging problem, particularly in crowded scenes with occlusion, appearance ambiguity, and non-linear motion. Conventional MOT frameworks often rely on appearance-based Re-Identification (Re-ID) and Intersection-over-Union (IoU) of object bounding boxes for object association. However, these cues become unreliable when objects are visually similar or overlapping, and computing pixel-level IoU for segmentation masks can be computationally expensive.

In this thesis, we propose two complementary MOT frameworks that incorporate monocular depth and segmentation cues to improve robustness in association. The first zero-shot depth-aware framework is training-free and introduces a Hierarchical Alignment Score (HAS), a novel metric that combines coarse bounding box IoU with fine-grained mask-level IoU using promptable segmentation. This hierarchical formulation improves matching precision in cluttered or occluded scenes.

The second framework avoids computing segmentation IoU altogether. Instead, it leverages a self-supervised encoder to fuse and refine depth-segmentation features into temporally stable embeddings, which are then used as an additional similarity signal in the association process. This reduces computational overhead while improving robustness to noise and appearance variation.

Both approaches operate under the efficient Tracking-by-Detection (TBD) paradigm and extend conventional 2D association strategies with spatially expressive cues. Evaluations on DanceTrack and SportsMOT benchmarks with non-linear motion demonstrate competitive performance, highlighting the utility of depth and segmentation as underutilized, yet powerful, cues for robust non-linear MOT.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Maria Amer and Dr. Charalambos Poullis, for their exceptional guidance, insightful feedback, and unwavering support throughout the course of this research. Their mentorship has been instrumental in shaping both the technical depth and the academic rigor of this thesis. I am especially grateful for their patience and encouragement, which helped me navigate the challenges and stay motivated.

I would also like to thank the members of the High-Performance Computing (HPC) Facility Speed for providing reliable computational resources and timely technical assistance. Their support played a crucial role in enabling uninterrupted experimentation during this research.

In addition, I extend my heartfelt thanks to my family for their unconditional love, encouragement, and understanding. Their constant support provided me with the strength and stability to pursue and complete this journey. This work would not have been possible without their sacrifices and belief in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Multi Object Tracking

Multi-Object Tracking (MOT) plays a crucial role in various real-world applications, including autonomous driving, human-computer interaction, and intelligent surveillance systems. The primary objective of MOT is to maintain the identity of multiple targets across a video sequence, using video frames as input. This becomes more challenging in dynamic scenes due to frequent occlusions, cluttered environments, and rapid motion, especially when targets are visually similar or undergo appearance changes over time.

In this thesis, non-linear motion refers to object trajectories that change direction, speed, or orientation in unpredictable ways. This is especially common in scenarios such as dance performances and sports, where objects move freely without following predetermined paths. This behavior contrasts with the more linear motion patterns observed in pedestrian tracking, where subjects typically move steadily across frames with fewer abrupt changes.

Modern MOT frameworks predominantly follow one of two paradigms: *Tracking-by-Detection (TBD)* and *Joint Detection-ReIdentification (JDR)*. In the TBD paradigm, objects are first detected in individual frames, and then an association algorithm links these detections across time to form consistent trajectories. This modular design enables the use of state-of-the-art object detectors but heavily depends on reliable detections and strong association cues [1–3]. Alternatively, JDR approaches aim to learn both object detection and identity embedding in a unified, end-to-end trainable

framework [4, 5]. These methods exploit shared feature representations to simultaneously localize objects and compute appearance features for identity matching. While JDR approaches reduce hand-crafted engineering between detection and tracking modules, they still struggle under linear motion prediction and object similarity due to their reliance on 2D visual features (e.g., appearance cues).

With the advent of deep learning models, particularly those based on Convolutional Neural Networks (CNNs) and Transformer architectures, substantial progress has been achieved [4–6]. These models have demonstrated competitive performance on public benchmarks and have even been adopted in industrial-grade systems. However, challenges such as identity switches (IDSW), detection noise, and long-term occlusions still persist. Occlusion, in particular, remains one of the most persistent bottlenecks [7]. It occurs when one or more targets are partially or fully hidden due to scene geometry, object overlap, or camera viewpoint. This causes degradation in both the association and Re-Identification (Re-ID) components of MOT pipelines, often resulting in drift or tracking failure [8, 9].

Two primary strategies have emerged to address occlusion: (1) improving association strategies by introducing stronger temporal cues, such as through sequence-level reasoning [10–12], and (2) leveraging auxiliary signals like depth [13, 14] or memory modules [12] to maintain identity continuity across missing observations. While these approaches are promising, many are constrained by reliance on synthetic training data or simplistic assumptions that fail under real-world complexity.

Therefore, there is a need for MOT models that are robust to occlusion, capable of maintaining identity across challenging temporal gaps, and generalizable to varied motion patterns and scene contexts. This thesis is motivated by these limitations, aiming to develop an MOT framework that improves robustness under occlusion and enhances identity consistency by leveraging more spatial-temporal cues.

1.2 Datasets

To evaluate our framework's effectiveness across a variety of challenging tracking environments, we conduct experiments on DanceTrack [15], SportsMOT [16], MOT17 [17], and MOT20 [18].

Each dataset presents distinct scenarios.

1.2.1 DanceTrack:

This dataset introduces challenging tracking conditions, where individuals move in non-linear patterns due to performance-based motion, and face frequent occlusions and crossovers. DanceTrack is structured with 40 training, 25 validation, and 35 testing sequences, emphasizing appearance similarity and spatial overlap. This dataset's complexity in non-linear motion and close-proximity interactions makes it ideal for assessing performance in dynamic environments.

1.2.2 SportsMOT:

Like DanceTrack, SportsMOT involves dynamic, non-linear motions, capturing fast and unpredictable subject movements typical in sports scenarios. Objects frequently interact and occlude each other in rapid sequences, making appearance and motion cues critical for consistent tracking. SportsMOT allows us to test the framework's adaptability to scenarios with distinct motion patterns and close-range activity.

1.2.3 MOT17 and MOT20:

Both MOT17 and MOT20 are established benchmarks for pedestrian tracking, with relatively linear motion patterns compared to DanceTrack and SportsMOT. MOT17 captures various urban scenes with moderate density and occasional occlusions, while MOT20 is densely crowded, featuring significant overlap and occlusion among pedestrians. These datasets are essential for evaluating tracking performance in crowded, urban scenes with a focus on linear trajectories. However, their emphasis on tracking objects with temporally consistent orientations with respect to the camera may limit the direct applicability of our framework, which is optimized for dynamic, non-linear motion with temporally arbitrary orientations.

1.3 Evaluation Metrics and Notation

Multi-Object Tracking (MOT) metrics provide a quantitative basis for evaluating tracking performance in terms of detection accuracy, identity preservation, and computational efficiency. These metrics are grounded in fundamental counting variables, such as True Positives (TP), False Positives (FP), False Negatives (FN), and Identity Switch (IDSW), which are typically computed at each frame and aggregated over time. Let t denote the time or frame index in a video sequence. We distinguish between detection-level quantities, denoted with the subscript D, which reflect errors in object presence (irrespective of identity), and association-level quantities, denoted with the subscript A, which reflect errors in maintaining identity continuity across frames (e.g., TP_D for detection-level true positives and TP_A for correctly associated tracks).

1.3.1 Higher Order Tracking Accuracy (HOTA)

HOTA is designed to equally evaluate both detection and association performance. It is computed as:

$$\text{HOTA} = \sqrt{\text{DetA} \cdot \text{AssA}}, \tag{1}$$

where DetA is detection accuracy, and AssA is association accuracy.

1.3.2 Association Accuracy (AssA)

AssA quantifies how well the tracker maintains correct identity matches for already-detected objects:

$$\text{AssA} = \frac{TP_A}{TP_A + FP_A + FN_A}, \tag{2}$$

where TP_A is the correct identity matches, FP_A is the incorrect identity matches, and FN_A is the missed identity associations.

1.3.3 ID-based F1 Score (IDF1)

IDF1 evaluates identity preservation by computing the harmonic mean of identity precision and identity recall:

$$\text{IDF1} = \frac{2 \cdot IDTP}{2 \cdot IDTP + IDFP + IDFN}, \tag{3}$$

where $IDTP$ is the number of correctly identified detections, $IDFP$ is the number of false identity assignments, and $IDFN$ is the number of missed identity matches. IDF1 is computed over the entire sequence to assess global identity consistency across time.

1.3.4 Multi-Object Tracking Accuracy (MOTA)

MOTA combines detection and identity-switch errors. It is computed as:

$$\text{MOTA} = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t}, \tag{4}$$

where FN_t is false negatives at frame t, FP_t is false positives at frame t, $IDSW_t$ is identity switches at frame t, and GT_t is Total ground truth objects at time t.

Although MOTA includes identity switches, the dominant terms in the numerator are often FN and FP, which are pure detection errors. As reported in [19], identity switches typically occur less frequently than missed or spurious detections. Therefore, a tracker with strong detection but poor identity consistency can still achieve a high MOTA score. This causes MOTA to be biased more toward detection quality than association performance.

1.3.5 Detection Accuracy (DetA)

DetA measures how accurately objects are detected, independent of their identities:

$$\text{DetA} = \frac{TP_D}{TP_D + FP_D + FN_D}, \tag{5}$$

where TP_D is correctly detected objects, FP_D is false detections, and FN_D is missed detections.

1.3.6 Frames Per Second (FPS)

FPS measures the speed of the tracking algorithm:

$$\text{FPS} = \frac{N}{T}, \tag{6}$$

where N is number of video frames processed, T is total processing time (seconds).

Among these metrics, HOTA, IDF1, and AssA provide deeper insights into identity preservation, while MOTA and DetA emphasize detection quality. FPS captures runtime efficiency.

1.4 Thesis Statement

The field of Multi-Object Tracking (MOT) has witnessed substantial improvements with the integration of deep learning models. However, despite advances in Tracking-by-Detection [1–3, 6, 20] and Joint Detection-ReIdentification frameworks [4, 5], the persistent challenge of maintaining consistent object identities in crowded, occluded, and dynamically changing scenes remains unsolved. Specifically, identity switches (IDSW), missed associations, and trajectory fragmentation persist as critical issues in benchmark evaluations [21].

There is an ongoing discussion in the MOT community regarding two key directions for improvement: enhancing model architectures and enabling more effective multi-frame (temporal) data association. Although many works focus on improving model architectures, such as through attention mechanisms [6, 12] or by modeling sequence-level context using graph-based reasoning [10, 11], these methods often fall short under real-world occlusions and long-term disocclusion intervals.

This thesis addresses the persistent problem of degraded tracking performance under occlusion and identity confusion. Our goal is to develop an MOT framework that (i) better preserves identity consistency across severe occlusions and motion clutter, (ii) is robust to noisy detections and partial observability, and (iii) can generalize to diverse scenes with non-linear object dynamics. To this end, we propose novel models that leverage enriched spatial-temporal cues with depth and segmentation masks and advanced association strategies to overcome the limitations of Tracking-by-Detection

frame-wise models. Thus, we seek to bridge this gap through methods that move beyond conventional 2D reasoning by incorporating rich geometric cues into the MOT pipeline.

1.5 Summary of Contributions

This thesis aims to address key challenges in Multi-Object Tracking (MOT), particularly in scenarios characterized by non-linear motion, frequent occlusions, and appearance ambiguity. We propose a novel, training-free framework that integrates spatial-temporal cues with a novel alignment strategy. In addition, we present a complementary method that leverages self-supervised depth-segmentation encoded features for enhanced association. The primary contributions of these works are summarized as follows:

- **Depth-aware Object Association:** We introduce the first MOT frameworks that explicitly incorporate monocular depth estimation as an independent cue during data association. By leveraging zero-shot depth maps, our methods enable robust spatial differentiation between overlapping or interacting objects.
- **Hierarchical Alignment Score (HAS):** We propose a novel alignment score that combines bounding box Intersection-over-Union (IoU) with pixel-level mask similarity. Unlike standard mask-IoU approaches that rely solely on geometric overlap, our method incorporates semantic similarity for robust matching.
- **Training-free Generalization and Robust Performance:** One of our proposed methods operates without task-specific training or fine-tuning. Nevertheless, it achieves competitive or state-of-the-art results on non-linear motion benchmarks such as DanceTrack and SportsMOT, particularly excelling in association metrics including HOTA, IDF1, and AssA.
- **Self-supervised Encoder for Feature Refinement:** We design a self-supervised encoder that refines noisy depth-segmentation embeddings, enhancing temporal stability and spatial discriminability while reducing the computational complexity of pixel-level similarity computations.

The codebases for both frameworks are publicly available:

- Depth-Aware Scoring and Hierarchical Alignment:
  https://github.com/Milad-Khanchi/DepthMOT
- Self-Supervised Depth and Mask-Aware Association:
  https://github.com/Milad-Khanchi/SelfTrEncMOT

1.6 Thesis Outline

This thesis is composed of four chapters, with Chapters 2 and 3 based on two original research manuscripts. These chapters present distinct, yet complementary, depth-aware frameworks for Multi-Object Tracking (MOT).

- **Chapter 2** is based on our accepted paper at the IEEE ICIP 2025, titled *Depth-Aware Scoring and Hierarchical Alignment for Multiple Object Tracking* [22]. This chapter proposes a training-free MOT framework that leverages zero-shot monocular depth estimation and introduces a novel Hierarchical Alignment Score (HAS) to combine bounding-box and pixel-level similarity. This method demonstrates generalization across occluded and crowded scenes without dataset-specific fine-tuning.
- **Chapter 3** is based on our paper under-review at the BMVC 2025. It presents a complementary approach titled *Fast Self-Supervised Depth and Mask-Aware Association for Multi-Object Tracking*. This framework avoids computing explicit mask IoU by introducing a self-supervised encoder that fuses depth and segmentation features into temporally consistent embeddings. These embeddings serve as an additional similarity signal, enhancing robustness and computational efficiency.
- **Chapter 4** concludes the thesis by summarizing key findings and discussing potential directions for future research, particularly toward making MOT more scalable and generalizable using geometry-aware representations.

Chapter 2

Depth-Aware Scoring and Hierarchical Alignment For Multiple Object Tracking

2.1 Introduction

Multiple object tracking (MOT) [1, 20, 23] involves detecting objects in video frames and continuously tracking them across time, requiring the simultaneous resolution of object detection, data association, and trajectory prediction. Challenges include noisy observations, occlusion, rapid motion, or similar objects. Recent research has focused on hybrid models that integrate both motion and appearance-based features. However, even with appearance cues, these approaches remain limited in scenarios that require spatial differentiation. For instance, two visually similar objects at different distances from the camera may appear indistinguishable in both motion and appearance, leading to frequent tracking errors. Moreover, occlusion scenarios often disrupt trajectory continuity, as traditional 2D Intersection over Union (IoU) based association is ineffective when objects overlap or move along similar paths. Furthermore, appearance-based approaches only extract appearance from cropped bounding boxes, neglecting the spatial context within each frame.

To address these challenges, we propose a novel depth-aware MOT framework that integrates monocular depth estimation into the tracking pipeline, allowing our framework to leverage 3D spatial cues for more robust association. By introducing zero-shot depth estimation, our framework

differentiates between objects based on their distance from the camera, providing a strong discriminative feature for association. Our depth-aware tracking model is further enhanced by a new Hierarchical Alignment Score (HAS) for association, which combines bounding box IoU with pixel-level alignment, capturing both coarse and fine object alignment features to improve matching accuracy. Unlike conventional IoU-based approaches, HAS adaptively refines the alignment between objects by emphasizing shape similarity in scenarios where bounding box overlap alone may be insufficient. Our HAS is hierarchical because it progressively combines coarse spatial alignment (via bounding box IoU) with fine-grained pixel-level similarity (via segmentation IoU) in a sequential manner. This layered matching formulation mimics a top-down refinement process: initial matches are scored based on geometric overlap, and only then enhanced by detailed shape consistency. Although not iterative in the optimization sense, this structure imposes a hierarchy in the way spatial and semantic alignment are composed, aligning with how object association evolves across frames.

To our knowledge, this is the first MOT framework to integrate monocular depth as an independent factor in the object association process for MOT. Our contributions are:

- We introduce a depth-aware tracking framework that incorporates zero-shot monocular depth estimation, providing robust spatial differentiation between objects based on their distance from the camera, thus improving association in complex scenarios.
- We propose a Hierarchical Alignment Score (HAS), a novel score that combines bounding box IoU with pixel-level alignment, enabling our framework to achieve precise object matching in cluttered and occluded environments.
- We present a comprehensive evaluation of our framework on challenging MOT benchmarks, showing that our approach resolves ambiguities in scenarios, such as occlusions and visually similar objects, without any training nor fine-tuning.

2.2 Related Works

MOT approaches have evolved from simple IoU-based matching to more sophisticated models that incorporate appearance features, motion cues, and graph-based techniques. MOT approaches

can be broadly categorized into two paradigms: joint detection-ReIdentification (JDR) [4, 5], and Tracking-by-Detection (TBD) [1–3, 6, 20].

Joint Detection-ReIdentification (JDR): JDR frameworks streamline MOT by combining object detection and association within an end-to-end model [4, 5]. For example, FairMOT [4] employs a parallel-branch architecture for simultaneous detection and Re-ID, improving robustness in crowded scenes. TransCenter [5] leverages a deformable Transformer to enhance detection and Re-ID across overlapping instances, achieving efficient network performance in complex scenes.

Tracking-by-Detection (TBD): Most high-performing MOT methods operate within a TBD paradigm, where objects are detected in individual frames before being linked across time [1–3, 6, 20, 24]. These models rely on consistent detection quality and effective data association strategies, enabling strong tracking performance across benchmarks.

Tracking can be performed in a frame-wise manner [1–3, 6], where the model processes each frame sequentially, utilizing information from previous frames. Alternatively, it can be conducted at the sequence level [7, 10, 11], where the entire sequence of frames is available during the association process.

2.2.1 Sequence-Level Tracking

Tracking by Segmentation: MOT and segmentation models have recently gained popularity, leveraging segmentation cues to improve association accuracy. For example, Brasó et al. [7] proposed a graph-based structure for segmentation and tracking, enhancing robustness under occlusions.

Tracking by Graph: Graph-based tracking methods represent detected entities as nodes and use edges to capture associations. In works by Brasó et al. [10] and Cetintas et al. [11], entities across frames are represented as nodes, with edge classification guiding associations. Hierarchical graphs, as seen in Cetintas et al. [11], process frame pairs in layers, iteratively refining associations across the sequence.

Self-Supervised Tracking: Lu et al. [25] propose path consistency as a reliable self-supervised signal to learn a robust object-matching model for MOT.

2.2.2 Frame-Wise Tracking

Tracking by Attention: Attention-based methods leverage the Transformer architecture to associate objects. Zhu et al. [24] introduced a dual-matching attention network to match detected objects, while TrackFormer [6] integrates detection and tracking in an end-to-end attention-based model, further refined to MOTRv2 by Zhang et al. [26] with YOLOX detector [27] for improved association. Gao et al. [12] implemented an end-to-end long-term memory-augmented transformer for MOT to refine detection and association.

Tracking by Regression: Many frame-wise tracking models apply regression-based methods to associate detected objects across frames [1–3, 23]. Kalman Filter-based approaches, such as OC-SORT [3], primarily leverage motion cues, while advanced models like Deep OC-SORT [2] and DiffMOT [1] enhance the association process with appearance cues to improve tracking in dense scenes. Huang et al. [28] addressed confusion issues in MOT by proposing an association method and detection post-processing technique.

Tracking by Depth: Quach et al. [13] incorporate dynamic control variables into a Kalman Filter, updating object states based on relative depth—defined as the depth ordering of detected objects w.r.t. the camera. However, their evaluation focused on the accuracy of depth estimation against RGB-D ground truth without demonstrating its direct impact on tracking performance. Wang et al. [14] utilize depth from stereo cameras alongside camera pose data. This approach relies heavily on accurate camera intrinsics and pose calibration. Liu et al. [29] propose a depth-aware tracking model, yet limited to indoor tracking.

Our work builds on appearance and motion-based models by integrating depth as a novel spatial feature as an independent decision matrix in the association step. In a frame-wise manner, our approach resolves ambiguities in challenging scenarios, such as occlusions and visually similar objects, without requiring training our motion and depth predictor on each dataset.

2.3 Methodology

In a TBD paradigm, objects are initially detected in individual frames and subsequently associated across time in a frame-wise manner. For object detection, we use the known model

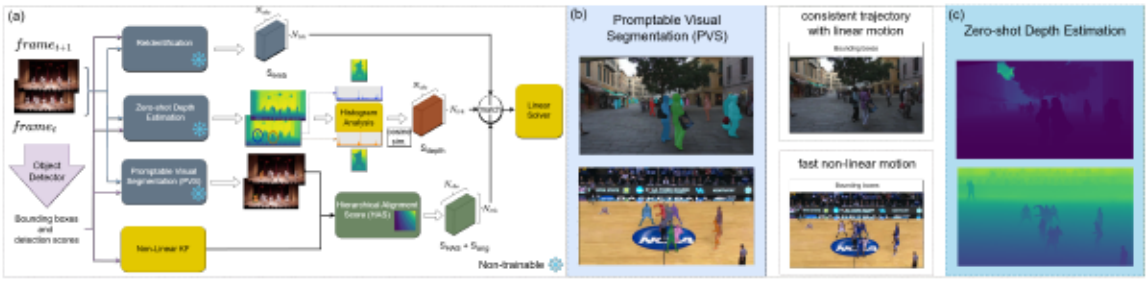

Figure 2.1: (a): Overview of the proposed framework, which integrates appearance scores from ReIdentification (RE-ID), motion scores derived from HAS, and depth scores. We use an advanced linear solver module and incorporate a PVS module for precise motion analysis. The Histogram Analysis block, as well as the RE-ID and HAS blocks, generate individual score matrices $R \in \mathbb{R}^{N_{obs} \times N_{trk}}$, where N_{obs} represents the number of new observations in the current frame and N_{trk} denotes the number of tracklets. These score matrices capture similarities based on appearance, motion, and depth, enabling a comprehensive assessment for object association. The Histogram Analysis block, in particular, also performs a comparison of pixel intensity distributions between the depth maps of two objects within the same frame, highlighting variations in depth based on their distances from the camera. (b), (c): Examples. Zero-shot monocular depth estimation and PVS modules in two scenarios.

YOLOX [30]. We perform object association in two steps. In the first step, we categorize detected objects based on their detection confidence (DF) as in [31]: high-confidence ($DF > 0.6$) and low-confidence ($DF \leq 0.6$). We match high-confidence objects with existing tracklets from previous frames using our matching technique detailed in Sec. 2.3.5. In the second step, the remaining unmatched objects are associated with future positions of remaining tracklets, solely based on the 2D bounding box overlap (IoU). We predict the position of future tracklets using a non-linear Kalman filter (see Sec. 2.3.1).

Current MOT frameworks heavily rely on IoU and 2D appearance-based ReIdentification (RE-ID) models, which may fail under heavy occlusion or similar objects. To address these limitations, our framework consists of four main components, as illustrated in Fig. 2.1 (a): an appearance-motion fusion model (Sec. 2.3.1), a depth-aware association process (Sec. 2.3.3), and two modules for refining motion-based object associations — Promptable Visual Segmentation (PVS) (Sec. 2.3.2) and a novel Hierarchical Alignment Score (HAS) (Sec. 2.3.4). Notably, we avoid training or fine-tuning any components, relying on existing pre-trained models to ensure the generalization and adaptability of our framework across various scenarios.

2.3.1 Appearance-Motion Fusion for MOT

Non-linear Kalman Filter: Our DepthMOT framework incorporates non-linear Kalman Filter (KF) to create motion scores. KF-based models operate in two main steps: prediction and update.

In the **Prediction Step**, the state estimate x at time t is derived based on the state estimate at $t-1$ using the transition model:

$$\begin{aligned} \hat{x}_{t|t-1} &= \mathbf{F}_t \hat{x}_{t-1|t-1}, \\ \mathbf{P}_{t|t-1} &= \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t^\top + \mathbf{Q}_t, \end{aligned} \tag{7}$$

where $\mathbf{F}_t$ is the state transition matrix, representing state evolution over time. $\mathbf{P}_{t|t-1}$ is the error covariance matrix, captures the uncertainty of the predicted state. $\mathbf{P}_{t|t-1}$ is updated by incorporating the process noise covariance matrix $\mathbf{Q}_t$, accounting for model uncertainties.

In the **Update Step**, the Kalman Gain $\mathbf{K}_t$ is computed to balance the predicted state against the observation as follows:

$$\begin{aligned} \mathbf{K}_t &= \mathbf{P}_{t|t-1} \mathbf{H}_t^\top \left(\mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^\top + \mathbf{R}_t \right)^{-1}, \\ \hat{x}_{t|t} &= \hat{x}_{t|t-1} + \mathbf{K}_t (z_t - \mathbf{H}_t \hat{x}_{t|t-1}), \\ \mathbf{P}_{t|t} &= (\mathbf{I} - \mathbf{K}_t \mathbf{H}_t) \mathbf{P}_{t|t-1}, \end{aligned} \tag{8}$$

where $\mathbf{P}_{t|t-1}$ is the error covariance matrix, and $\mathbf{R}_t$ is the observation noise covariance. $\mathbf{K}_t$ weights the correction based on the observed measurement z_t. This leads to an updated state estimate and reduced uncertainty in $\mathbf{P}_{t|t}$. OC-SORT [3] introduces Observation-Centric Re-Update (ORU) to enhance KF robustness and address scenarios with lost detections. ORU compensates for missed observations by re-updating KF parameters as follows:

$$\begin{aligned} \mathbf{K}_t &= \mathbf{P}_{t|t-1} \mathbf{H}_t^\top \left(\mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^\top + \mathbf{R}_t \right)^{-1}, \\ \hat{x}_{t|t} &= \hat{x}_{t|t-1} + \mathbf{K}_t (\tilde{z}_t - \mathbf{H}_t \hat{x}_{t|t-1}), \\ \mathbf{P}_{t|t} &= (\mathbf{I} - \mathbf{K}_t \mathbf{H}_t) \mathbf{P}_{t|t-1}, \end{aligned} \tag{9}$$

where $\tilde{\mathbf{z}}_t$ is the trajectory interpolating between the last-seen observation $\tilde{\mathbf{z}}_{t_1}$ and the re-associated observation $\tilde{\mathbf{z}}_{t_2}$ using:

$$\tilde{\mathbf{z}}_t = \text{Traj}_{virt}(\tilde{\mathbf{z}}_{t_1}, \tilde{\mathbf{z}}_{t_2}, t), \quad t_1 < t < t_2 \tag{10}$$

where Traj_{virt} denotes the interpolation function, enabling backtracking over missing frames.

We define each KF state as $x = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T$, where (u, v) denote the 2D coordinates of the object's center, s the bounding box scale, and r the aspect ratio of the bounding box (assumed constant). The derivatives $\dot{u}$, $\dot{v}$, and $\dot{s}$ correspond to the temporal changes in u, v, and s, respectively [3]. Additionally, each detected object provides a bounding box defined as $d = [u, v, w, h, c]^T$, where (u, v) are the center coordinates, w and h are the width and height, and c the detection confidence.

As shown in Eq. 13, in the association step, our framework calculates a matching score matrix linking observations in the current frame to existing tracklets. This matrix includes two motion-based components: S_{IoU} and S_{ang}. The score S_{IoU} is based on the IoU between bounding boxes, which measures their spatial overlap. In parallel, S_{ang} captures the directional similarity between the new observations and existing tracklets. Higher values for S_{IoU} and S_{ang} reflect greater similarity between the observation and the corresponding tracklet, while lower values suggest divergence [3].

ReIdentification: To extract appearance features, we utilize the pre-trained FastReID [32], a RE-ID model based on convolutional neural networks (CNN) that has been trained for the MOT [1, 2]. FastReID extracts features from each cropped object in the frame, which are then compared across tracklets and observations using cosine similarity, yielding an appearance-based score matrix S_{emb}. To maintain consistent appearance features for each tracklet, we apply an exponential moving average (EMA) [33, 34], updating each tracklet's embeddings after each association step.

The embedding update for each tracklet is given by:

$$emb_t = \mathcal{C} \cdot emb_{t-1} + (1 - \mathcal{C}) \cdot emb_{new}, \tag{11}$$

where $\mathcal{C}$ is a dynamic coefficient defined as in [2], $\mathcal{C} = \mathcal{T} + (1 - \mathcal{T}) \cdot \left(1 - \frac{c - \text{thresh}}{1 - \text{thresh}}\right)$. In this formulation, c represents the detection confidence, thresh is a detection threshold, and $\mathcal{T}$ is a fixed parameter, set to 0.95 following prior work.

2.3.2 Promptable Visual Segmentation (PVS)

For each frame, the object detection outputs bounding boxes. To have a fine object shape alignment, we apply Promptable Visual Segmentation (PVS), which extends static image segmentation

to the video domain by enabling the generation of a spatio-temporal mask for a segment of interest across frames. In PVS, a prompt, such as a point, bounding box, or initial mask, is applied to a frame to define the target object, and the model then propagates this mask through subsequent frames. This task has emerged as a powerful method for fine-grained object segmentation in video sequences [35–37].

For shape alignment, we integrate Segment Anything Model 2 (SAM2) [38], an advanced PVS framework designed for images and videos. SAM2 has some tracking capabilities but is not designed to track objects by matching bounding boxes and ID continuity, as required in MOT. Our shape alignment approach works as follows: 1) Given the bounding boxes of tracklets (i.e., objects already being tracked) in the previous frame, SAM2 generates segmentation masks for these tracklets. These masks represent the precise object shape within the bounding boxes. 2) For objects newly detected in the current frame t, SAM2 uses the bounding box of each detected object in frame t to propagate the segmentation backward to find the corresponding object mask in frame $t-1$. This allows us to retrieve the segmentation of each newly detected object in the previous frame, even if that object was not explicitly tracked before. 3) Once we have segmentation masks for both the existing tracklets (from frame $t-1$) and the newly detected objects (propagated backward from frame t), we perform mask matching within frame $t-1$ by computing the mask IoU between each tracklet's mask and the mask of each newly detected object. More details on how these IoUs are integrated into our association process can be found in Sec. 2.3.4. By incorporating this segmentation-based mask matching, our method ensures that newly detected objects in frame t are correctly linked to their corresponding tracklets from frame $t-1$, thereby improving tracking consistency and reducing ID switches.

2.3.3 Zero-Shot Depth Estimation

To enrich our MOT framework with 3D spatial context, we leverage zero-shot depth estimation to capture object distances from the camera. Conventional zero-shot depth models have often relied on camera-specific metadata, such as intrinsics, which can be unavailable or unreliable. Recent methods, such as ZeroDepth [39] and DMD [40], perform well in predicting depth maps. However, these methods are limited by their dependence on precise camera parameters, which reduces

generalizability. Depth Pro [41] overcomes this limitation by achieving accurate depth estimation without requiring intrinsics. With its ability to estimate depth at high speed and accuracy, as well as to predict focal length from a single image, Depth Pro is ideally suited for MOT applications where efficiency and generalization are critical.

We process each frame in the video sequence using Depth Pro, which produces a dense depth map of the scene. This map provides a relative spatial representation, allowing us to measure the depth of each object by isolating the depth values within its bounding box. Depth-based measurements add a valuable discriminative layer by distinguishing visually similar objects that occupy different spatial planes, improving tracking accuracy in crowded scenarios.

For frame-to-frame comparison of depth maps, we evaluated two approaches detailed in the ablation study (Sec. 2.5.5). We found that histogram-based vectorizations of depth values yielded the most effective results. By constructing a histogram of depth values within each object's bounding box, we obtain a compact yet powerful descriptor of the object's 3D characteristics. To determine the depth similarity between frames, we compute the cosine similarity between these histograms, quantifying the similarity of spatial distributions.

As shown in Eq. 13, we integrate the resulting depth similarity score into the overall matching score matrix. Fig. 2.1 (a) illustrates this approach with two objects in the same frame, where histogram differences highlight the depth variance corresponding to their distance from the camera. This example demonstrates the added discriminatory power depth brings to object tracking. Fig. 2.1 (b) and (c) shows examples of zero-shot depth estimation and PVS modules.

2.3.4 Hierarchical Alignment Score (HAS)

IoU is widely used in MOT to measure spatial overlap and associate objects across consecutive frames. Traditional IoU measures the overlap between bounding boxes and ranges from 0 to 1, capturing spatial alignment but overlooking important shape-based cues. Consequently, purely motion-based methods often fall short in complex scenes with occlusions, where bounding boxes alone may not accurately represent object shape or position. While recent appearance-based approaches [2, 33] incorporate visual features, these methods only extract appearance from cropped bounding boxes, neglecting the spatial context within each frame.

To address these limitations, we introduce a novel matching metric, HAS, which combines coarse alignment via bounding box IoU with fine object shape alignment via pixel-level IoU, progressively refining object matching. HAS enhances the matching robustness by integrating both spatial and shape information, ensuring that matches are not only spatially aligned but also consistent in shape alignment.

We calculate the HAS score S_{HAS} as follows:

$$S_{\text{HAS}}(\hat{X}, D) = S_{\text{IoU}_{bbox}}(\hat{X}, D) \times exp^{S_{\text{IoU}_{Seg}}(\hat{X}, D)}, \tag{12}$$

where $\hat{X}$ denotes the predicted state of a tracklet, including its bounding box or segmentation mask estimated in the current frame, and D represents a new detection in the current frame, including the observed bounding box or segmentation mask. $S_{\text{IoU}_{bbox}}$ represents the bounding box IoU, facilitating initial coarse matching based on location and scale, while $S_{\text{IoU}_{Seg}}$ denotes the pixel-wise IoU, capturing precise shape alignment between objects. The exponential weighting on $S_{\text{IoU}_{Seg}}$ ensures that minor improvements in shape alignment lead to significant increases in the HAS score, reinforcing high-fidelity matches. While in Eq. 12 we use an exponential weighting to emphasize fine-grained shape alignment, other non-linearities such as quadratic or sigmoid functions could be explored in future work to modulate the influence of segmentation-based similarity. Each non-linearity presents a different rate of amplification and saturation behavior, which may offer advantages under different noise levels or object density conditions.

The core innovation of HAS lies in its hierarchical alignment process. During the initial stages, the matching function prioritizes bounding box IoU, yielding a coarse spatial alignment based on object location and scale. As bounding box alignment improves, the influence of the exponential pixel-wise IoU term $exp^{S_{\text{IoU}_{Seg}}}$ becomes more pronounced, refining alignment based on the actual object shapes. This hierarchical refinement ensures robust association by balancing location and shape information, progressively emphasizing fine shape details as alignment quality increases. Fig. 2.2 illustrates a heatmap of S_{HAS}, showing how it evolves through the initial stages as a function of $S_{\text{IoU}_{bbox}}$ and $S_{\text{IoU}_{Seg}}$; the exponential weighting sharply amplifies the influence of fine-grained segmentation early on, particularly when bounding box overlap is low, emphasizing precise matching.

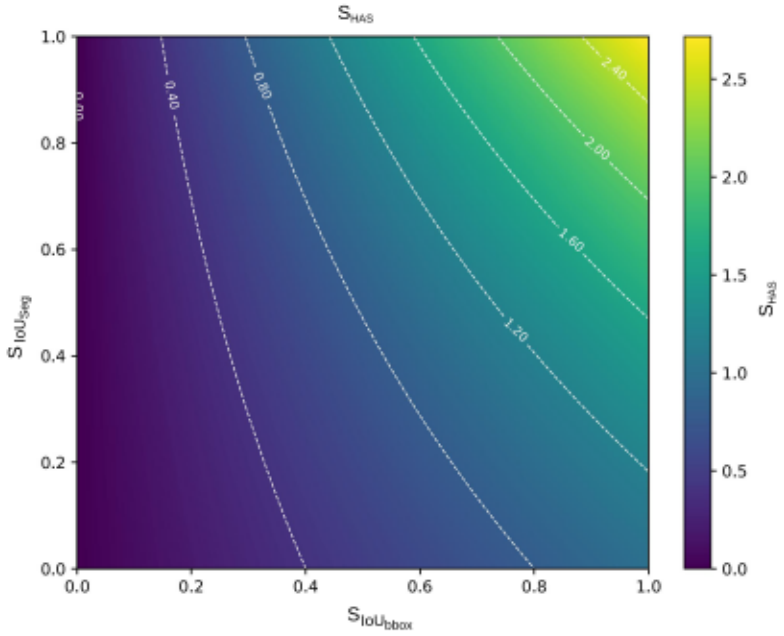

Figure 2.2: Hierarchical Alignment Score (HAS). Heatmap of S_{HAS} as a function of bounding box IoU ($S_{\text{IoU}_{bbox}}$) and segment IoU ($S_{\text{IoU}_{Seg}}$). Contours indicate score levels, demonstrating the hierarchical influence of both spatial and shape alignment on the overall score. A high S_{HAS} suggests high similarity, while a low S_{HAS} indicate dissimilarity

2.3.5 Total Matching Score and Linear Solver

To further enhance tracking performance, we incorporate HAS into a comprehensive score function, which integrates motion, appearance, and depth cues as follows:

$$\begin{aligned} Match_t = S_{\text{HAS}_t}(\hat{X}, D) + S_{\text{ang}_t}(\hat{X}, D) \\ + S_{\text{depth}_t}(\hat{X}, D) + S_{\text{emb}_t}(\hat{X}, D), \end{aligned} \tag{13}$$

where $S_{\text{emb}_t}(\hat{X}, D)$ the appearance score, $S_{\text{depth}_t}(\hat{X}, D)$ depth similarity, $S_{\text{ang}_t}(\hat{X}, D)$ captures motion direction alignment [2], and $S_{\text{HAS}_t}(\hat{X}, D)$ the proposed HAS score. This holistic approach creates a robust multi-cue framework that effectively balances motion, 2D and 3D spatial alignment, and appearance features.

The matching score matrix $Match_t$ in Eq. 13 is then negated, transforming it into a cost matrix compatible with the linear solver [42] for optimal assignment.

2.4 Experimental Results

2.4.1 Datasets and Evaluation Metrics

Datasets. To evaluate our framework's effectiveness across a variety of challenging tracking environments, we conduct experiments on DanceTrack, SportsMOT, MOT17, and MOT20. Each dataset presents distinct scenarios.

DanceTrack: This dataset introduces challenging tracking conditions, where individuals move in non-linear patterns due to performance-based motion, and face frequent occlusions and crossovers. DanceTrack is structured with 40 training, 25 validation, and 35 testing sequences, emphasizing appearance similarity and spatial overlap. This dataset's complexity in non-linear motion and close-proximity interactions makes it ideal for assessing performance in dynamic environments.

SportsMOT: Like DanceTrack, SportsMOT involves dynamic, non-linear motions, capturing fast and unpredictable subject movements typical in sports scenarios. Objects frequently interact and occlude each other in rapid sequences, making appearance and motion cues critical for consistent tracking. SportsMOT allows us to test the framework's adaptability to scenarios with distinct motion patterns and close-range activity.

MOT17 and MOT20: Both MOT17 and MOT20 are established benchmarks for pedestrian tracking, with relatively linear motion patterns compared to DanceTrack and SportsMOT. MOT17 captures various urban scenes with moderate density and occasional occlusions, while MOT20 is densely crowded, featuring significant overlap and occlusion among pedestrians. These datasets are essential for evaluating tracking performance in crowded, urban scenes with a focus on linear trajectories. However, their emphasis on tracking objects with temporally consistent orientations with respect to the camera may limit the direct applicability of our framework, which is optimized for dynamic, non-linear motion with temporally arbitrary orientations.

Metrics. Main performance metrics for MOT are Higher Order Tracking Accuracy (HOTA) [19], Association Accuracy (AssA), and ID-based F1 Score (IDF1) [43]. HOTA assesses both detection and association accuracy. IDF1 and AssA primarily evaluate association performance. Multi-Object Tracking Accuracy (MOTA) [44] and Detection Accuracy (DetA) metrics focus on detection accuracy. Regarding computational efficiency, frames per second (FPS) is the primary metric reported in

most MOT studies. It is typically measured for the tracking component (i.e., the association stage), excluding the computational cost of feature extraction, such as detection and ReID [3].

2.4.2 Implementation Details

Model. To ensure consistency and fair comparison with recent studies [1–3], we adopt YOLOX [30] as the object detector. In our tables, MOT methods using YOLOX as the detector are highlighted in blue and bold indicates the best performance among TBD trackers (bottom part of each table).

Inference. For inference, all experiments are conducted on a single NVIDIA A100 GPU with a batch size of one. Our method do not require any training. We note that comparing FPS across different MOT models is not plausible due to discrepancies in hardware and software configurations of each model. For instance, [45] uses an NVIDIA GeForce RTX 3090, while [46] employs an NVIDIA Tesla V100. On the DanceTrack validation set, our association technique achieves an inference speed of 3.35 FPS. The primary computational bottleneck lies in the computation of mask IoUs. We adhere to the standard tracking hyperparameters (e.g., detection thresholds and update smoothing factors) from [2] to ensure consistent evaluation of the association component across methods.

2.4.3 Benchmark Evaluation

DanceTrack: The performance of DepthMOT on DanceTrack is shown in Table 2.1. Compared to the state-of-the-art TBD trackers, like DiffMOT [1], Deep OC-SORT [2], and CMTrack [23], DepthMOT achieves superior results without training for specific datasets. Our framework improves HOTA by 1.97%, AssA by 3.84%, and IDF1 by 3.54% over DiffMOT. Both AssA and IDF1 indicate the association performance and highlight the effectiveness of using our HAS and depth. The state-of-the-art JDR tracker MOTRv2 [26] is designed for non-linear motion and outperforms our method for this dataset; however, our TBD method is more efficient. MOTRv2 demands 8 Tesla V100 GPUs for multi-stage training. Also, our method outperforms MOTRv2 under linear motion datasets MOT17 and MOT20 (see Table 2.3).

SportsMOT: Compared to the state-of-the-art DiffMOT [1], our DepthMOT has overall comparable performance on the SportsMOT dataset, as shown in Table 2.2. It outperforms in IDF1 and AssA,

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
FairMOT [4]	39.7	40.8	23.8	82.2	66.7
CenterTrack [47]	41.8	35.7	22.6	86.8	78.1
TraDes [48]	43.3	41.2	25.4	86.2	74.5
TransTrack [49]	45.5	45.2	27.5	85.4	75.9
DiffusionTrack [45]	52.4	47.5	33.5	89.5	82.2
MOTR [46]	54.2	51.5	40.2	79.7	73.5
MeMOTR [12]	68.5	71.2	58.4	89.9	80.5
MOTRv2 [26]	69.9	71.7	59.0	91.9	83.0
GHOST [50]	56.7	57.7	39.8	91.3	81.1
DeepSORT [51]	45.6	47.9	29.7	87.8	71.0
ByteTrack [31]	47.3	52.5	31.4	89.5	71.6
SORT [42]	47.9	50.8	31.2	91.8	72.0
MotionTrack [52]	52.9	53.8	34.7	91.3	80.9
OC-SORT [3]	55.1	54.2	38.0	89.4	80.3
StrongSORT++ [34]	55.6	55.2	38.6	91.1	80.7
GeneralTrack [53]	59.2	59.7	42.8	91.8	82.0
C-BIoU [54]	60.6	61.6	45.4	91.8	81.3
Deep OC-SORT [2]	61.3	61.5	45.8	92.3	82.2
CMTrack [23]	61.8	63.3	46.4	92.5	-
DiffMOT [1]	62.3	63.0	47.2	**92.8**	**82.5**
DepthMOT	**64.27**	**66.54**	**51.04**	90.08	81.07

Table 2.1: Comparison with MOT trackers on the DanceTrack **test set**. Methods are grouped into JDR in the upper part, followed by TBD in the lower part. Our DepthMOT outperforms the state-of-the-art TBD tracker DiffMOT, which requires training on each dataset separately.

which are critical for evaluating identity preservation and association accuracy.

MOT17 and MOT20: We evaluate DepthMOT on MOT17 and MOT20 under the *private* detection protocol (i.e., using method-specific object detections). As shown in Table 2.3, none of the trackers achieve state-of-the-art performance regarding metrics in MOT17 and MOT20. However, DepthMOT records the lowest false positive (FP) rate of 1.3. This reduced FP rate highlights the effectiveness of HAS, which minimizes false associations (grouping different objects in the same tracklets) and fragmentation.

Analysis of the results: Based on Tab. 2.1 and Tab. 2.2, our framework with hierarchical alignment

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
FairMOT [4]	49.3	53.5	34.7	86.4	70.2
GTR [55]	54.5	55.8	45.9	67.9	64.8
QDTrack [56]	60.4	62.3	47.2	90.1	77.5
CenterTrack [47]	62.7	60.0	48.0	90.8	81.7
TransTrack [49]	68.9	71.5	57.5	92.6	82.7
ByteTrack [31]	62.8	69.8	51.2	94.1	77.1
BoT-SORT [33]	68.7	70.0	55.9	94.5	84.4
OC-SORT [3]	71.9	72.2	59.8	94.5	86.4
*ByteTrack [31]	64.1	71.4	52.3	95.9	78.5
*MixSort-Byte [16]	65.7	74.4	58.4	96.2	78.8
*OC-SORT [3]	73.7	74.0	61.5	96.5	88.5
*MixSort-OC [16]	74.1	74.4	62.0	96.5	88.5
*GeneralTrack [53]	74.1	76.4	61.7	96.8	89.0
*DiffMOT [1]	**76.2**	76.1	65.1	**97.1**	**89.3**
*DepthMOT	**76.2**	**76.9**	**65.6**	95.9	88.5

Table 2.2: Comparison of JDR and TBD trackers on the SportsMOT **test set**. * indicates that the detector is trained on the SportsMOT train and validation sets. Our method outperforms in IDF1 and AssA, which reflect how consistently a tracker preserves object identities and maintains accurate associations over time.

scoring and depth-aware association performs well in scenarios like DanceTrack and SportsMOT, where rapid, non-linear motions and complex inter-object interactions are prevalent. As shown in Table 2.3, no single model achieves superior results across metrics in the MOT17 and MOT20 benchmarks. However, linear motion prediction models, such as CMTrack [23] and Deep OC-SORT [2], demonstrate better performance. This can be due to the consistent and predictable pedestrian trajectories in these datasets. Additionally, as illustrated in Fig. 2.3 and Fig. 2.1 (c), zero-shot depth estimation faces challenges in outdoor environments with low illumination, producing low-contrast depth maps that struggle to distinguish distant objects effectively. This limitation impacts the accuracy of DepthMOT on MOT17 and MOT20. Despite these factors, DepthMOT, with no training in individual datasets, performs well in complex, high-occlusion environments while also demonstrating adaptability across a variety of tracking scenarios.

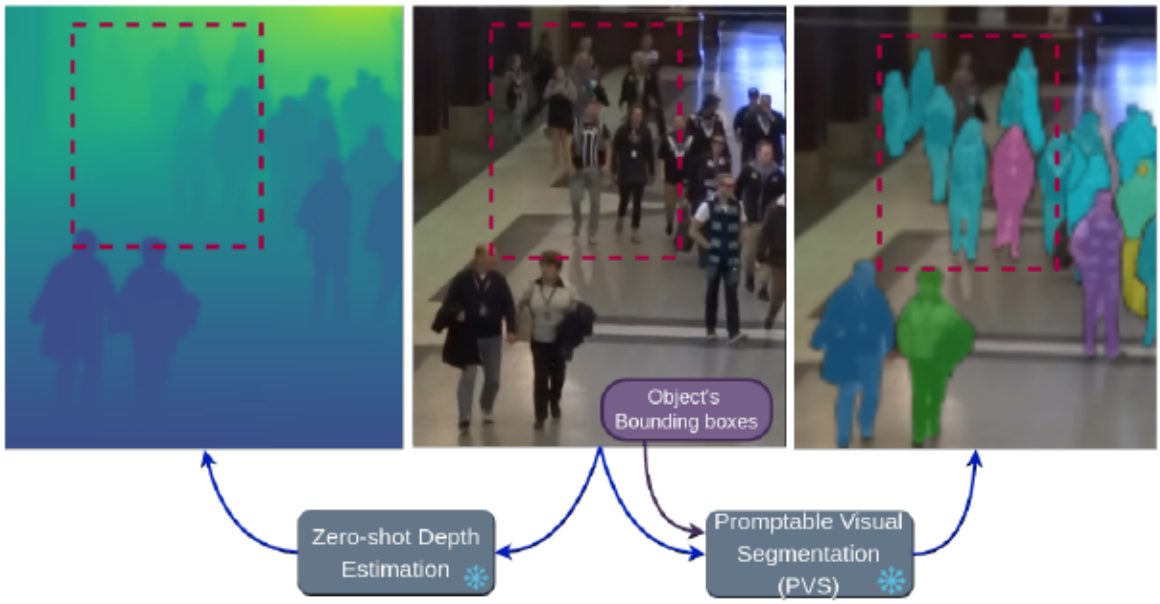

Figure 2.3: **Challenges.** An example of zero-shot depth estimation and PVS modules, emphasizing the encountered challenges under different lighting conditions in the MOT20 dataset. The highlighted area, marked by a dotted square, illustrates that the depth map of certain objects is not accurately predicted.

MOT17									
Tracker	**HOTA↑**	**IDF1↑**	**AssA↑**	**MOTA↑**	**FP(10^4)↓**	**FN(10^4)↓**	**IDs↓**	**Frag↓**	**AssR↑**
FairMOT [4]	59.3	72.3	58.0	73.7	2.75	11.7	3303	8073	63.6
TransCT [5]	54.5	62.2	49.7	73.2	2.31	12.4	4614	9519	54.2
TransTrack [49]	54.1	63.5	47.9	75.2	5.02	8.64	3603	4872	57.1
CSTrack [21]	59.3	72.3	-	74.9	2.38	11.4	3567	-	-
QDTrack [56]	53.9	66.3	52.7	68.7	2.66	14.66	3378	8091	57.2
MOTR [46]	57.2	68.4	55.8	71.9	2.11	13.6	2115	3897	59.2
CenterTrack [47]	52.2	64.7	-	67.8	1.8	1.6	3039	-	-
MeMOTR [12]	58.8	71.5	58.4	72.8	-	-	-	-	-
DiffusionTrack [45]	60.8	73.8	58.8	77.9	-	-	3819	4815	-
TransMOT [57]	61.7	75.1	59.9	76.7	3.62	9.32	2346	7719	66.5
MixSort-OC [16]	63.4	77.8	63.2	78.9	-	-	1509	-	-
MixSort-Byte [16]	64.0	78.7	64.2	78.9	-	-	2235	-	-
C-BIoU [54]	64.1	79.7	63.7	81.1	-	-	-	-	-
MOTRv2 [26]	62.0	75.0	60.6	78.6	-	-	-	-	-
GHOST [50]	62.8	77.1	-	78.7	-	-	2325	-	-
ByteTrack [31]	63.1	77.3	62.0	80.3	2.55	8.37	2196	2277	68.2
DeconfuseTrack [28]	64.9	80.6	65.1	80.4	-	-	-	-	-
MotionTrack [52]	65.1	80.1	-	**81.1**	2.38	**8.16**	1140	-	-
OC-SORT [3]	63.2	77.5	63.2	78.0	1.51	10.8	1950	2040	67.5
StrongSORT [34]	63.5	78.5	63.7	78.3	-	-	1446	-	-
StrongSORT++ [34]	64.4	79.5	64.4	79.6	2.79	8.62	1194	1866	**71.0**
GeneralTrack [53]	64.0	78.3	63.1	80.6	-	-	1563	-	-
Deep OC-SORT [2]	64.9	80.6	65.9	79.4	1.66	9.88	1023	2196	70.1
CMTrack [23]	**65.5**	**81.5**	**66.1**	80.7	2.59	8.19	**912**	**1653**	-
DiffMOT [1]	64.5	79.3	64.6	79.8	-	-	-	-	-
DepthMOT	62.7	77.9	63.6	76.5	**1.3**	11.7	1342	-	68.8

MOT20									
Tracker	**HOTA↑**	**IDF1↑**	**AssA↑**	**MOTA↑**	**FP(10^4)↓**	**FN(10^4)↓**	**IDs↓**	**Frag↓**	**AssR↑**
GSDT [58]	53.6	67.5	52.7	67.1	3.19	13.5	3,131	9,875	58.5
CSTrack [21]	54.0	68.6	54.0	66.6	2.54	14.4	3,196	7,632	57.6
FairMOT [4]	54.6	67.3	54.7	61.8	10.3	8.89	5,243	7,874	60.7
DiffusionTrack [45]	55.3	66.3	51.3	72.8	-	-	4,117	4,446	-
TransMOT [57]	61.9	75.2	60.1	77.5	3.42	8.08	1,615	2,421	66.3
MOTRv2 [26]	60.3	72.2	58.1	76.2	-	-	-	-	-
GHOST [50]	61.2	75.2	-	73.7	-	-	1,264	-	-
ByteTrack [31]	61.3	75.2	59.6	77.8	2.62	8.76	1,223	1,460	66.2
DeconfuseTrack [28]	63.3	77.6	62.7	**78.1**	-	-	-	-	-
MotionTrack [52]	62.8	76.5	61.8	78.0	2.86	**8.41**	1,165	1,321	-
OC-SORT [3]	62.1	75.9	62.0	75.5	1.80	10.8	913	1,198	67.5
StrongSORT [34]	61.5	75.9	63.2	72.2	-	-	1,066	-	-
StrongSORT++ [34]	62.6	77.0	64.0	73.8	1.66	11.8	770	1,003	69.6
GeneralTrack [53]	61.4	74.0	59.5	77.2	-	-	1,627	-	-
Deep OC-SORT [2]	63.9	79.2	65.7	75.6	1.69	10.8	779	1,536	**70.8**
CMTrack [23]	**64.8**	**79.9**	**66.7**	76.2	2.22	10.04	**730**	**987**	-
DiffMOT [1]	61.7	74.9	60.5	76.7	-	-	-	-	-
DepthMOT	62.4	77.3	64.3	73.2	**1.3**	12	1,141	-	68.6

Table 2.3: Results on MOT17-test and MOT20-test. The lower parts show TBD methods, which are relevant to ours. Methods in the blue blocks use the same YOLOX detector. As can be seen, no tracker performs best across metrics and datasets. Our DepthMOT has the lowest false positive (FP).

2.5 Ablations

To evaluate the impact of two key components: the proposed Hierarchical Alignment Score (HAS) and the integration of depth information, we conduct a comprehensive ablation study on MOT17 and the DanceTrack validation sets, which represent linear and non-linear motion patterns, respectively.

2.5.1 Effect of HAS in association:

We first evaluate whether traditional IoU metrics (bounding box IoU or pixel-wise IoU) can independently have the performance of the proposed HAS method. Table 2.4 shows that using either bounding box IoU or pixel-wise IoU alone results in suboptimal performance on DanceTrack, as they fail to adequately capture the nuanced alignment between object contours and spatial position. HAS, by combining both metrics in a hierarchical structure, provides a more balanced approach that aligns objects more accurately across frames.

2.5.2 Effect of Depth in association:

To examine the contribution of depth information, we compare tracking performance with and without depth scores. Table 2.4 presents different framework configurations with or without depth scores and the HAS technique. As shown, on DanceTrack, the combination of both HAS and depth information yields the highest accuracy, achieving 61.81% HOTA, 49.12% AssA, and 64.13% IDF1. This improvement demonstrates that depth is an effective auxiliary feature, enhancing object association by providing an additional spatial cue that helps distinguish objects in scenes with high overlap or occlusion.

					DanceTrack-val			**MOT17-val**		
Appearance	Mask IoU	Bbox IoU	HAS	Depth	HOTA ↑	AssA ↑	IDF1 ↑	HOTA ↑	AssA ↑	IDF1 ↑
✓	✓				54.78	38.54	52.70	68.22	66.81	77.2
✓		✓			59.46	43.92	59.08	**70.42**	**70.83**	**80.73**
✓			✓		60.62	47.15	62.16	69.85	69.65	79.71
✓		✓		✓	60.45	47.02	62.13	70.19	70.65	80.52
✓			✓	✓	**61.81**	**49.12**	**64.13**	69.88	69.73	79.7

Table 2.4: Ablation study: Impact of HAS and depth on DanceTrack and MOT17 **validation set** performance.

2.5.3 HAS in First and Second Matching Steps:

We first evaluate the effect of applying the HAS technique at different stages of the matching process. As shown in Table 2.5, there is no difference in the performance metrics for both experiments. To enhance the model's generalizability, we applied the HAS technique only in the first step for all datasets.

	DanceTrack-val			MOT17-val		
HAS Technique	HOTA ↑	AssA ↑	IDF1 ↑	HOTA ↑	AssA ↑	IDF1 ↑
First Step Only	61.81	49.12	64.13	**69.88**	**69.73**	**79.7**
First & Second Steps	**61.82**	**49.15**	**64.15**	69.79	69.57	79.57

Table 2.5: **Ablation.** Comparison of tracking performance with different stages of the HAS technique applied during the matching process on the DanceTrack validation set. "First Step Only" applies HAS to the first matching Step, while "First & Second Steps" applies HAS to both primary and secondary matching steps.

2.5.4 Depth score in First and Second Matching step:

We further investigate the effect of applying depth scores in the first and second matching steps. Table 2.6 reveals that using depth in the first matching step yields the highest performance, while applying depth in the second step does not contribute positively. Table 2.5 and 2.6 align with the observations made in [2], where the authors, noting the absence of appearance scores in their model for the second round of matching, have inferred that appearance features are not helpful for the second step.

	DanceTrack-val			MOT17-val		
Depth Score	HOTA ↑	AssA ↑	IDF1 ↑	HOTA ↑	AssA ↑	IDF1 ↑
First Step Only	**61.81**	**49.12**	**64.13**	**69.88**	**69.73**	**79.7**
First & Second Steps	61.69	48.94	63.66	69.87	**69.73**	79.66

Table 2.6: **Ablation.** Evaluation of incorporating depth in the first and second matching steps on the DanceTrack validation set.

2.5.5 Kernel Density Estimation (KDE) vs. Histogram:

We conducted an experiment to assess the impact of various vectorization techniques on the final depth score. Kernel Density Estimation (KDE) is a well-established method for converting pixel values into a probability distribution. As demonstrated in Table 2.7, the histogram approach exhibits superior performance compared to KDE across all metrics in both datasets. This enhanced performance of the histogram method may be attributed to its effective representation of depth distributions within object bounding boxes. In contrast to KDE, which tends to smooth over local variations, histograms accurately capture the frequency distribution of depth values without averaging, thereby avoiding the potential over-smoothing of significant spatial variations. In scenarios where precise distinction of depth at close range is essential, histograms may more effectively preserve critical differences, thereby improving depth-informed association accuracy in complex scenes.

	DanceTrack-val			**MOT17-val**		
Depth Vectorization Method	HOTA ↑	AssA ↑	IDF1 ↑	HOTA ↑	AssA ↑	IDF1 ↑
KDE	61.26	48.35	62.84	69.53	69.35	79.32
Histogram	**61.81**	**49.12**	**64.13**	**69.88**	**69.73**	**79.7**

Table 2.7: **Ablation.** Evaluation of different vectorization techniques for converting depth maps into similarity scores on the DanceTrack and MOT17 validation set.

2.6 Visual Results

Fig. 2.4 and Fig. 2.5 illustrate the tracking performance of our DepthMOT framework on challenging sequences from the DanceTrack and SportsMOT datasets, respectively. Additionally, Fig. 2.6 and Fig.2.7 showcase the results of DepthMOT on MOT17 and MOT20. For enhanced visualization, bounding boxes of individuals with similar IDs are displayed in similar colors, with the unique ID indicated in red at the top of each bounding box. However, some colors may appear visually similar due to the high density of individuals in specific frames.

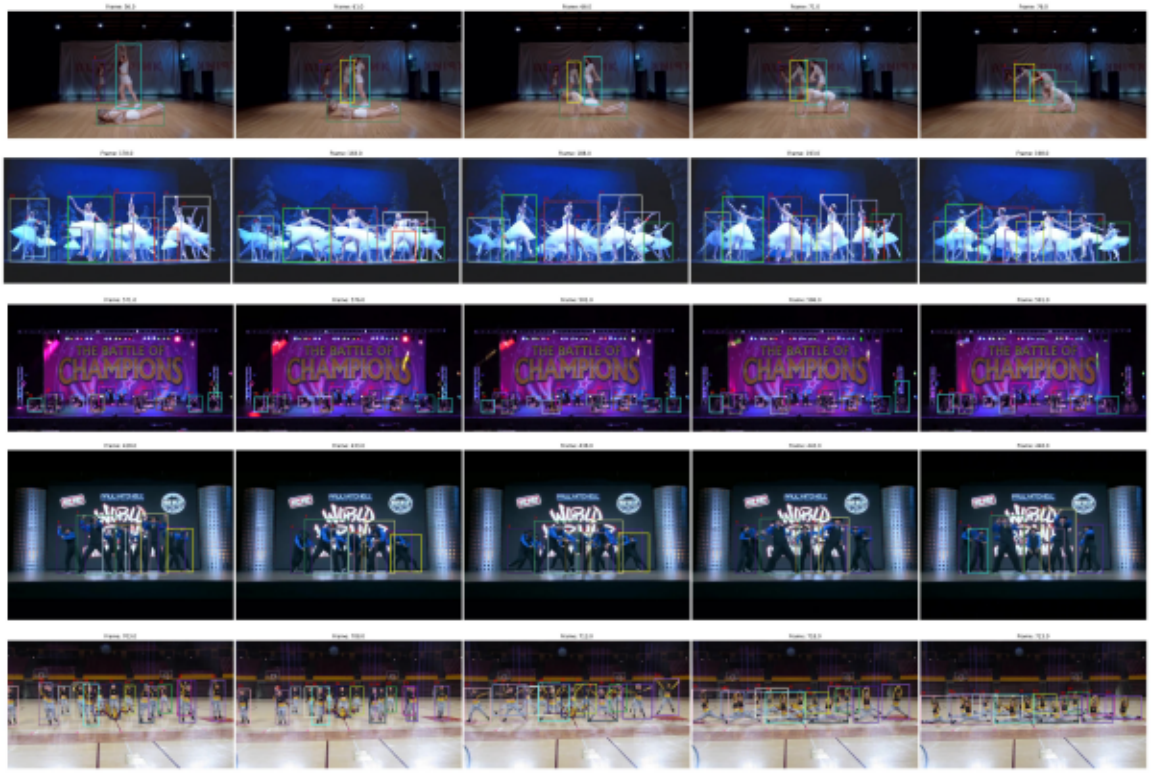

Figure 2.4: **Examples.** DanceTrack tracking results

Figure 2.5: **Examples.** SportsMOT tracking results

Figure 2.6: **Examples.** MOT17 tracking results

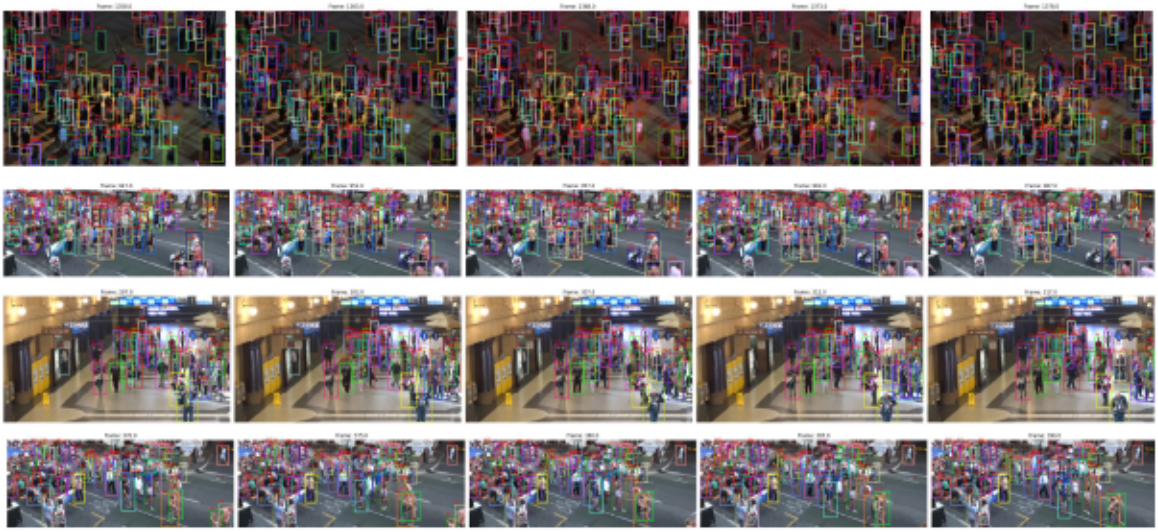

Figure 2.7: **Examples.** MOT20 tracking results

2.7 Summary

We proposed the DepthMOT framework for MOT, which incorporates depth information alongside visual and motion features. Additionally, we introduced a hierarchical alignment score (HAS)

that prioritizes bounding boxes during the initial association phase and progressively balances pixel-level segments as matching improves. DepthMOT achieves state-of-the-art results on DanceTrack and SportsMOT, handling rapid non-linear object motion, high appearance similarity, frequent occlusions, and crossovers. Notably, DepthMOT achieves competitive performance using only pre-trained models without any training nor fine-tuning.

Chapter 3

Fast Self-Supervised depth and mask aware Association for Multi-Object Tracking

3.1 Introduction

Multi-Object Tracking (MOT) [1, 23, 59] aims to detect and maintain object identities across video frames. Despite notable advances, existing approaches still struggle under conditions of occlusion, appearance similarity, and rapid motion. These challenges are amplified when 2D cues, such as bounding box overlap (IoU) and appearance re-identification (Re-ID), are the sole drivers of association. For instance, two pedestrians walking in parallel but at different depths may appear indistinguishable in 2D, leading to frequent ID switches.

To address these limitations, we propose a combined depth and segmentation aware method that supplements traditional 2D cues with pixel-aligned geometric reasoning. Specifically, we combine zero-shot monocular depth estimation with promptable segmentation masks to extract fine-grained spatial features. The combined depth-segmentation embeddings are processed by a lightweight, self-supervised encoder that enhances temporal consistency and reduces noise caused by artifacts from segmentation or depth map. The resulting features serve as an additional matching score

during data association, complementing motion and appearance 2D cues.

Unlike prior works [13, 14, 29] that use depth or segmentation as auxiliary inputs, our approach introduces depth-segmentation cues as an explicit similarity matrix for matching. This design enables more robust identity preservation, particularly in crowded, occluded, or visually ambiguous scenes.

We validate our method on the challenging MOT benchmarks DanceTrack and SportsMOT, which feature crowded and occluded objects with complex non-linear motion, and demonstrate consistent improvements in association-based metrics such as HOTA, IDF1, and AssA. Our method achieves competitive performance on benchmarks with simpler, mostly linear motion, such as MOT17. Our contributions are: 1) We design a self-supervised encoder to enhance depth-segmentation features' temporal stability and discriminability. 2) Our MOT method is the first to use the self-supervised encoder to refine segmentation masks and integrate them into the matching score without computing mask IoU. 3) We achieve competitive or state-of-the-art performance across various tracking scenarios, especially in dense and occluded scenes.

3.2 Prior Works

Joint Detection-ReID methods (JDR): They unify detection and tracking in a single forward pass [4, 5]. FairMOT [4] pioneered a dual-branch approach combining anchor-free detection with appearance embeddings. TransCenter [5] extends this with deformable attention, enabling improved occlusion handling. Recent models in this category go beyond earlier dual-branch architectures by integrating attention mechanisms or spatial alignment strategies. AFMTrack [60] enhances identity preservation by introducing an attention-based feature matching network, allowing robust association even in dense scenes. DilateTracker [61] integrates dilated attention modules into the joint detection-ReID framework, significantly boosting identity recall under occlusion.

Tracking by Detection methods (TBD): The tracking-by-detection paradigm remains dominant in recent literature, where objects are first localized in each frame and then associated temporally [1–3, 6, 20, 24, 26]. The effectiveness of these methods heavily relies on detection quality and the design of robust association strategies. Depending on their temporal processing technique, these methods

can be categorized into sequence-level and frame-wise tracking.

Sequence-Level Tracking: Graph-based approaches have gained popularity for maintaining long-term consistency. Brasó et al. [10] and Cetintas et al. [11] frame association as edge prediction in a spatio-temporal graph, while Lu et al. [25] introduce self-supervised learning via path consistency. However, these still rely primarily on appearance embeddings rather than spatial-aware cues.

Frame-Wise Tracking: Attention-based models have shown strong potential in object matching across frames. TrackFormer [6] and MOTRv2 [26] leverage transformer decoders for joint detection and tracking, while MeMOTR [12] incorporates long-term memory. They perform well under non-linear motion datasets such as DanceTrack but poorly under linear motion datasets such as MOT17. Earlier dual-attention designs [24] laid the foundation for such approaches. Regression-based methods like OC-SORT [3], Deep OC-SORT [2], and DiffMOT [1] emphasize motion continuity and efficient association. Confidence-aware [23] and post-correction strategies [28] have also emerged to reduce matching errors.

Depth-Aware and Self-Supervised Association: Several works have introduced depth into tracking. Quach et al. [13] apply relative depth ordering in Kalman filters, Wang et al. [14] combine stereo depth with pose estimation, and Liu et al. [29] propose a depth-aware tracker for indoor scenes. However, these methods treat depth as an auxiliary signal rather than a primary association cue. Self-supervised Re-ID learning has gained traction recently. Li et al. [59] embed self-supervision into FairMOT-style pipelines using path consistency, improving generalization. Nonetheless, it depends on contrastive or clustering objectives and does not utilize fused 3D spatial information.

3.3 Proposed Approach

Our approach follows the tracking-by-detection paradigm, where objects are first localized in each frame and then associated over time in a frame-wise manner. For detection, consistent with recent MOT benchmarks, we use YOLOX [30]. As shown in Fig. 3.1, our method is composed of three main modules: a depth-segmentation fusion module that combines depth and mask features

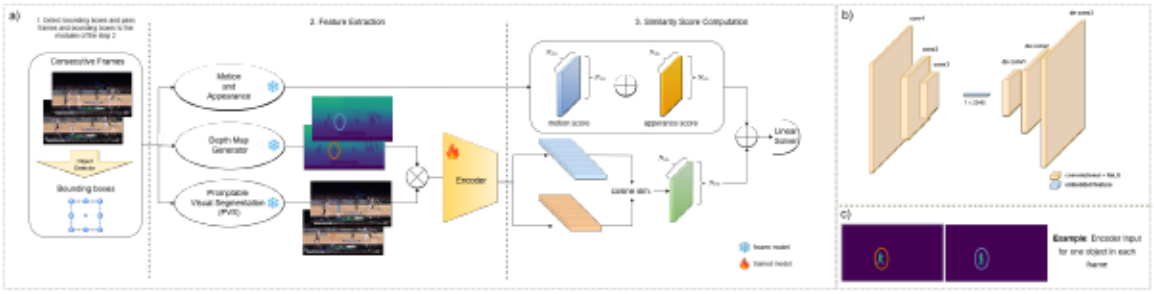

Figure 3.1: (a) Overview of SelfTrEncMOT. Given consecutive video frames and their object detector bounding boxes, we extract motion and appearance embeddings, and compute depth maps (via zero-shot monocular estimation) and segmentation masks (via Promptable Visual Segmentation). Depth and segmentation cues are fused into depth-segmentation embeddings and refined by a self-supervised encoder. The final association score integrates these embeddings with motion and appearance cues using a linear assignment solver. (b) Architecture of the depth-segmentation autoencoder. (c) Example of the encoder's input embedding.

for each object (See Sec. 3.3.1); a self-supervised encoder that refines these fused features into stable embeddings (See Sec. 3.3.2); and an appearance-motion module that extracts re-identification features and motion cues (See Sec. 3.3.3).

Together, these modules generate three similarity scores that are used during object association. Following ByteTrack [31], the association process is divided into two stages. In the first stage, high-confidence detections ($DC > 0.6$) are matched to existing tracklets using our multi-cue similarity scores (Sec. 3.3.4). In the second stage, unmatched detections are associated with tracklets using Intersection-over-Union (IoU), based on predicted positions from a non-linear Kalman filter.

3.3.1 Depth-Segmentation Fusion

Zero-Shot Depth Estimation: Each frame in the video sequence is processed using Depth Pro [41] to generate a depth map, which provides a relative spatial representation of the scene.

Promptable Visual Segmentation (PVS): To achieve fine-grained shape alignment beyond bounding boxes, we incorporate PVS, a method that enables segmentation mask propagation across frames by leveraging prompts (e.g., points, bounding boxes, or masks). PVS extends static image segmentation to the video domain, facilitating consistent mask generation for the same object across time [35–37].

For implementation, we use Segment Anything Model 2 (SAM2) [38], a prompt-driven segmentation framework designed for both images and videos. We use SAM2 for spatio-temporal shape alignment. Although SAM2 supports limited tracking, it is not designed for object ID consistency and bounding box matching as required in standard MOT tasks.

Our segmentation-driven fusion process proceeds as follows: For each tracklet in frame $t-1$, the corresponding bounding box is used as a prompt input to SAM2 to generate a precise segmentation mask that reflects object shape within the box. Next, for each newly detected object in frame t, its bounding box is used to prompt SAM2 in reverse (i.e., propagate backward) to recover its segmentation in frame $t-1$, aligning it with existing tracklets for direct association. Once segmentation masks are obtained for both existing tracklets (from $t-1$) and new detections (backward-propagated to $t-1$), we perform pixel-wise multiplication of each mask with its corresponding depth map. This yields fused depth-segmentation embeddings that encode both fine-grained object shape and relative spatial location.

3.3.2 Self-Supervised Depth-Segmentation Encoder

Computing the mask IoU between segmentation outputs is computationally expensive and introduces major latency into the tracking pipeline. Additionally, both segmentation and depth estimation used in the depth-segmentation fusion are subject to limitations. SAM2 may produce misaligned masks when propagating segments across frames, particularly under fast motion or occlusion. We use DepthPro [41], a recent zero-shot monocular depth estimator, to obtain dense depth maps from RGB frames. DepthPro, while effective, can generate noisy depth maps in regions with poor texture or challenging lighting. These imperfections degrade the stability and reliability of the fused features.

To address this, we introduce a lightweight depth-segmentation encoder designed specifically to suppress noise and enhance the temporal consistency of fused features. The encoder is part of the compact convolutional autoencoder that learns to denoise and compress the fused maps into discriminative embeddings suitable for tracking. The encoder uses three convolutional layers with kernel size 4×4 and stride 2, increasing channels from 1 to 32, 64, and 128, followed by batch normalization and ReLU activations. The resulting feature map is flattened and passed through a linear

layer to produce a 2048-dimensional bottleneck. The decoder mirrors this with transposed convolutions that progressively upsample and reduce channels back to 1. The bottleneck embeddings, computed per object, serve as compact descriptors for matching, while reconstruction ensures that the bottleneck retains key structural cues. This encoder is trained in a self-supervised manner to enhance the discriminative quality and temporal consistency of the fused depth-segmentation features. The training process of our autoencoder proceeds as follows:

(1) At each training step, the depth-segmentation embeddings of tracklets from frame $t-1$ and newly detected objects from frame t are passed through a shared autoencoder.

(2) A reconstruction loss is computed using Mean Squared Error (MSE) between each input and its reconstructed output:

$$\mathcal{L}_{\text{recon}} = \left\| f_i - \hat{f}_i \right\|_2^2, \tag{14}$$

where f_i is the original depth-segmentation map, $\hat{f}_i$ is its reconstruction.

(3) Additionally, we apply MSE loss between the bottleneck (latent) representations of the tracklet at frame $t-1$ and its corresponding detection at frame t:

$$\mathcal{L}_{\text{bottleneck}} = \|b_{t-1} - b_t\|_2^2, \tag{15}$$

where b_{t-1} and b_t are the bottleneck embeddings for frame $t-1$ and frame t, respectively.

(4) The final objective is a sum of the two losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{bottleneck}}. \tag{16}$$

This self-supervised refinement not only filters out noise from the segmentation and depth sources but also improves temporal coherence in the feature embeddings, leading to more reliable object association under challenging visual conditions. To further stabilize temporal dynamics in the encoder features across frames, inspired by [33, 34], we apply the tracklet embedding update strategy:

$$emb_t = \mathcal{C} \cdot emb_{t-1} + (1 - \mathcal{C}) \cdot emb_{new}, \tag{17}$$

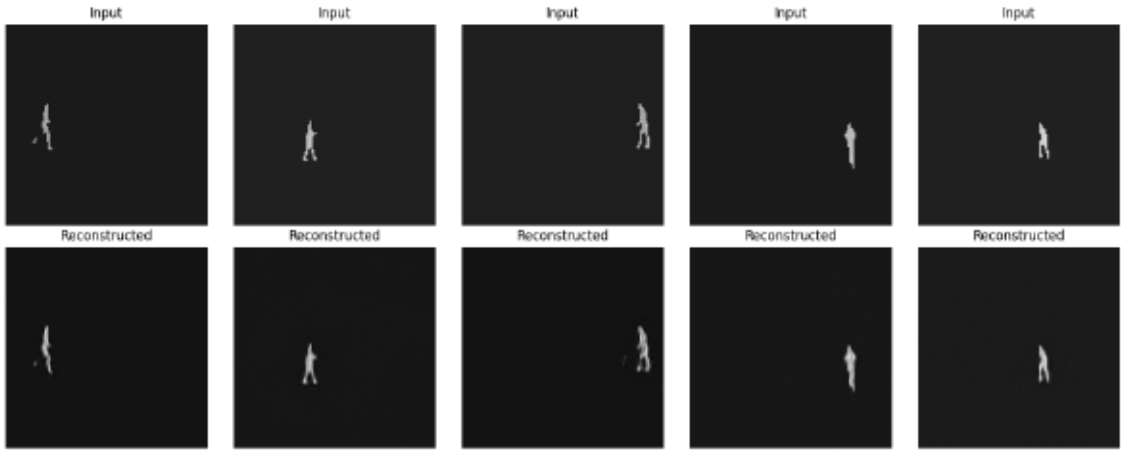

Figure 3.2: Qualitative results of the depth-segmentation autoencoder. Top: input fused embeddings; Bottom: reconstructions. The encoder preserves key spatial details and object boundaries, supporting robust association, despite variations in scale and structure.

where C is dynamically computed as in [2], $C = T + (1-T) \cdot \left(1 - \frac{DC-\text{thresh}}{1-\text{thresh}}\right)$, DC is the detection confidence score output by the object detector, thresh is a fixed confidence threshold (set to 0.6), and T is a hyperparameter (set to 0.95). Figure 3.2 represents the input-output pairs of the depth-segmentation autoencoder, where the top row shows the input fused depth and mask-aware maps, and the bottom row illustrates the corresponding reconstructions. This visualization highlights how the encoder preserves critical structural cues despite compression. This result shows that despite the major reduction in dimension, the reconstructed masks retain useful information. The refined embeddings output by the encoder are then compared using cosine similarity to compute the depth-segmentation similarity score S_{sd}. The raw depth-segmentation product is never directly used for matching; only the encoder-refined features participate in similarity computation.

3.3.3 Appearance-Motion Fusion for MOT

Nonlinear Kalman Filter: Our SelfTrEncMOT method uses a nonlinear Kalman Filter to model object motion dynamics and predict tracklet locations across frames. This filter operates in two main steps: it first propagates previous state estimates based on learned motion priors, then corrects predictions using new observations. To enhance robustness to missed detections, we incorporate the observation-centric re-update (ORU) mechanism from OC-SORT [3], which interpolates

virtual measurements when intermediate detections are absent. This approach allows tracklets to be updated more consistently over time, even during short-term occlusions. Each state includes bounding box geometry and motion velocities, enabling accurate position forecasting under complex dynamics.

Motion Matching: During association, a matching score matrix is computed using two motion-based components: S_{IoU} and S_{ang}. S_{IoU} measures spatial overlap between predicted and observed bounding boxes, while S_{ang} captures angular consistency in motion direction. Higher scores in both terms reflect stronger association confidence [3].

Appearance Matching: We adopt FastReID [32], a model based on convolutional neural networks (CNN) trained for MOT [1, 2], to extract the appearance embeddings from each detected object. These features are compared across frames using cosine similarity to compute an appearance-based score matrix S_{emb}.

3.3.4 Total Matching Score and Linear Solver

To associate detected bounding boxes with existing tracklets, we compute a total matching score that integrates motion, appearance, and depth-segmentation cues as follows:

$$Match_t = S_{\text{IoU}_t}(\hat{X}, D) + S_{\text{ang}_t}(\hat{X}, D) + S_{\text{sd}_t}(\hat{X}, D) + S_{\text{emb}_t}(\hat{X}, D), \tag{18}$$

where $S_{\text{IoU}_t}(\hat{X}, D)$ represents spatial overlap between predicted and observed bounding boxes, $S_{\text{ang}_t}(\hat{X}, D)$ captures angular motion similarity [2], $S_{\text{emb}_t}(\hat{X}, D)$ measures appearance similarity using cosine distance between FastReID embeddings, and $S_{\text{sd}_t}(\hat{X}, D)$ is the proposed depth-segmentation similarity score obtained from the cosine similarity between encoder embeddings. The matching score matrix $Match_t$ is then negated to form a cost matrix suitable for linear assignment. We use the Hungarian algorithm [42] to perform optimal data association at each time step.

3.4 Experimental Results

3.4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our approach across three diverse benchmarks to assess its effectiveness under varying motion dynamics, crowd densities, and interaction complexities.

SportsMOT features fast, unpredictable subject movement, common in athletic or competitive contexts. The motion is highly dynamic, and inter-object occlusions occur frequently and abruptly. This dataset tests the method's resilience to rapid motion shifts and closely interacting subjects, making both appearance and motion cues essential.

DanceTrack presents performance-driven scenarios where subjects exhibit highly non-linear motion, frequent occlusions, and close-proximity interactions. The dataset includes 40 training, 25 validation, and 35 test sequences, placing emphasis on appearance similarity and spatial ambiguity. These characteristics make DanceTrack particularly suitable for evaluating tracking robustness in dynamic and visually crowded environments.

MOT17 is a widely used benchmark for multi-pedestrian tracking in urban environments, featuring scenes of moderate crowd density and structured pedestrian movement. It serves as a standard for evaluating trackers under relatively linear and predictable motion. While useful for general-purpose benchmarking, MOT17 is less representative of the non-linear and appearance-ambiguous motion patterns targeted by our method.

Evaluation Metrics. For a comprehensive assessment, we adopt standard MOT evaluation metrics: Higher Order Tracking Accuracy (HOTA), Association Accuracy (AssA), Detection Accuracy (DetA) [19], Identification F1 Score (IDF1) [43], and Multi-Object Tracking Accuracy (MOTA) [44]. HOTA provides a balanced evaluation of detection and association accuracy. IDF1 and AssA specifically emphasize identity preservation and association quality, while MOTA focuses on detection-level performance. In terms of computational efficiency, we report frames per second (FPS) based on the tracking component, consistent with prior MOT works where runtime typically reflects the association stage [3].

3.4.2 Implementation Details

Model. For consistency and fair benchmarking, we adopt YOLOX [30] as the default object detector, following recent MOT methods [1–3]. In our tables, methods using YOLOX are marked in blue for clarity. The hyperparameter $\mathcal{T}$ is used in Eq. 17 to calculate C, and is set to 0.95. It is fixed for all datasets and experiments. We adhere to the standard tracking hyperparameters (e.g., detection thresholds and update smoothing factors) from [2] to ensure consistent evaluation of the association component across methods.

Training. SAM2 and DepthPro are not fine-tuned but used in their released form. The depth-segmentation autoencoder is trained offline, independently of the tracking method. Training is performed separately on fused depth-segmentation embeddings extracted from the training split of each dataset. We use a single NVIDIA A100 GPU with a batch size of 64 for all benchmarks. The method is trained for 12 epochs using the Adam optimizer with a learning rate of $1e^{-3}$. The loss function in Eq. 16 combines two Mean Squared Error (MSE) terms: one for reconstruction and another for bottleneck alignment across frames, as described in Section 3.3.2. To encourage stronger temporal consistency, only the bottleneck loss is used for the final two epochs. After training, only the encoder is retained and integrated into the tracker for online inference. We train our method for each dataset separately.

Inference. All experiments are conducted on a single NVIDIA A100 GPU with a batch size of one. Overall runtime of a MOT method (i.e., the feature extraction including object detection and the association parts) varies depending on the feature extractor backbones. As a result, FPS is often reported only for the association part [3]. Moreover, reported FPS values in prior works may not be directly comparable due to heterogeneous hardware setups (e.g., RTX 3090 in [45] vs. Tesla V100 in [46]). On the DanceTrack validation set, our tracker runs efficiently, with the association stage sustaining over 125 frames per second (FPS). In our results tables, we divide MOT methods into TBD (bottom part) and JDR (upper part), and mark in bold the best-performing method in each category.

3.4.3 Benchmark Evaluation

Results for SportsMOT: The SportsMOT benchmark poses unique challenges due to its dynamic, non-linear object trajectories, frequent occlusions, and visually similar targets interacting at close range. As shown in Table 3.1, compared to the state-of-the-art DiffMOT [1], our SelfTrEncMOT achieves better results across HOTA, AssA, and IDF1 association and identification metrics. These improvements reflect the effectiveness of integrating depth-segmentation cues and our self-supervised encoder. Together, these components support robust association in visually ambiguous and fast-paced sports tracking scenarios.

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
FairMOT [4]	49.3	53.5	34.7	86.4	70.2
CenterTrack [47]	62.7	60.0	48.0	90.8	81.7
TransTrack [49]	**68.9**	**71.5**	**57.5**	**92.6**	**82.7**
ByteTrack [31]	62.8	69.8	51.2	94.1	77.1
BoT-SORT [33]	68.7	70.0	55.9	94.5	84.4
OC-SORT [3]	71.9	72.2	59.8	94.5	86.4
DiffMOT [1]	72.1	72.8	60.5	94.5	86.0
*ByteTrack [31]	64.1	71.4	52.3	95.9	78.5
*MixSort-Byte [16]	65.7	74.4	58.4	96.2	78.8
*OC-SORT [3]	73.7	74.0	61.5	96.5	88.5
*MixSort-OC [16]	74.1	74.4	62.0	96.5	88.5
*GeneralTrack [53]	74.1	76.4	61.7	96.8	89.0
*DiffMOT [1]	76.2	76.1	65.1	**97.1**	**89.3**
*SelfTrEncMOT	**76.4**	**77.1**	**66.0**	95.84	88.4

Table 3.1: Comparison with MOT trackers on the SportsMOT **test set**. The lower part shows TBD methods, which are relevant to ours. Methods marked with * indicate the detector was trained on SportsMOT's train and validation sets. YOLOX-based methods are highlighted in blue. The top part includes JDR methods, which have higher computational requirements for training than TBD methods.

Results for DanceTrack: We evaluate SelfTrEncMOT on the DanceTrack benchmark in Table 3.2. Compared to the TBD state-of-the-art, such as DiffMOT [1], Deep OC-SORT [2], and CMTrack [23], our SelfTrEncMOT achieves better performance in the association metrics HOTA, IDF1, and AssA. The current state-of-the-art under this dataset is MOTRv2 [26], which is a Joint Detection-ReID, relying on a fully transformer-based architecture with joint training for detection and tracking. MOTRv2 requires significantly higher computational requirements than our method, e.g., 8x Tesla

V100 GPUs for multi-stage training. Our method requires 1x GPU and single-stage training. MOTRv2 underperforms for the linear motion dataset MOT17 (see Table 3.3).

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
FairMOT [4]	39.7	40.8	23.8	82.2	66.7
CenterTrack [47]	41.8	35.7	22.6	86.8	78.1
TransTrack [49]	45.5	45.2	27.5	85.4	75.9
MOTRv2 [26]	**69.9**	**71.7**	**59.0**	**91.9**	**83.0**
ByteTrack [31]	47.3	52.5	31.4	89.5	71.6
MotionTrack [52]	52.9	53.8	34.7	91.3	80.9
OC-SORT [3]	55.1	54.2	38.0	89.4	80.3
StrongSORT++ [34]	55.6	55.2	38.6	91.1	80.7
GeneralTrack [53]	59.2	59.7	42.8	91.8	82.0
C-BIoU [54]	60.6	61.6	45.4	91.8	81.3
Deep OC-SORT [2]	61.3	61.5	45.8	92.3	82.2
CMTrack [23]	61.8	63.3	46.4	92.5	-
DiffMOT [1]	62.3	63.0	47.2	**92.8**	**82.5**
SelfTrEncMOT	**64.14**	**66.47**	**50.85**	90.08	81.06

Table 3.2: Comparison on the DanceTrack **test set**. Methods are grouped into JDR in the upper part and TBD in the lower part. Methods using YOLOX are highlighted in blue. SelfTrEncMOT ranks first among TBD methods in identity association metrics (HOTA, AssA, IDF1). The state-of-the-art MOTRv2 requires significantly more computational resources than our method (e.g., 8x GPU versus 1x GPU for training).

Results for MOT17: As shown in Table 3.3, our method does not surpass state-of-the-art performance on MOT17. This is primarily because object motions in this dataset are mostly linear (e.g., cars and pedestrians moving in predictable directions). In such scenarios, motion patterns leave limited scope for depth information to provide substantial additional benefits. Nevertheless, SelfTrEncMOT achieves competitive results, demonstrating robust tracking performance comparable to leading methods.

3.4.4 Ablation Study

To assess the effectiveness of depth-segmentation guided association in our method, we perform an ablation study (Table 3.4) comparing three setups: (1) appearance with mask IoU, (2) appearance with bounding box IoU, and (3) appearance with bounding box IoU augmented by our depth-segmentation encoder. We evaluate on the *DanceTrack* and *MOT17* validation sets, which

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	FP(10^4)↓	FN(10^4)↓	ID_s↓	Frag↓	AssR↑
FairMOT [4]	59.3	72.3	58.0	73.7	2.75	11.7	3,303	8,073	63.6
CenterTrack [47]	52.2	64.7	-	67.8	1.8	1.6	3,039	-	-
TransTrack [49]	54.1	63.5	47.9	75.2	5.02	8.64	3,603	4,872	57.1
MOTRv2 [26]	**62.0**	**75.0**	**60.6**	**78.6**	-	-	-	-	-
ByteTrack [31]	63.1	77.3	62.0	80.3	2.55	8.37	2,196	2,277	68.2
MotionTrack [52]	65.1	80.1	-	**81.1**	2.38	**8.16**	1,140	-	-
OC-SORT [3]	63.2	77.5	63.2	78.0	**1.51**	10.8	1,950	2,040	67.5
StrongSORT++ [34]	64.4	79.5	64.4	79.6	2.79	8.62	1,194	1,866	**71.0**
GeneralTrack [53]	64.0	78.3	63.1	80.6	-	-	1,563	-	-
C-BIoU [54]	64.1	79.7	63.7	**81.1**	-	-	-	-	-
Deep OC-SORT [2]	64.9	80.6	65.9	79.4	1.66	9.88	1,023	2,196	70.1
CMTrack [23]	**65.5**	**81.5**	**66.1**	80.7	2.59	8.19	**912**	**1,653**	-
DiffMOT [1]	64.5	79.3	64.6	79.8	-	-	-	-	-
SelfTrEncMOT	63.48	78.12	63.25	79.16	1.9	9.6	1,008	-	68.8

Table 3.3: Comparison on the MOT17 **test set**. The JDR methods are in the upper part, and the TBD methods are in the lower part. Methods using YOLOX are highlighted in blue. State-of-the-art is CMTrack, which performs weakly for challenging datasets such as DanceTrack. MOTRv2 ranks lower than all TBD methods, including ours.

represent non-linear motion and structured pedestrian movement, respectively.

Ablation Setting				DanceTrack-val			MOT17-val		
Appearance	Mask IoU	Bbox IoU	Depth-Segmentation	HOTA↑	AssA↑	IDF1↑	HOTA↑	AssA↑	IDF1↑
✓	✓			54.78	38.52	52.71	68.26	66.81	77.20
✓		✓		59.46	43.93	59.11	70.43	70.83	80.73
✓		✓	✓	**60.61**	**47.04**	**62.34**	**72.22**	**71.79**	**82.52**

Table 3.4: Ablation study on DanceTrack and MOT17 **validation sets**. We evaluate combinations of appearance and IoU-based cues, with and without our encoder-refined depth-segmentation association module.

Complementarity with Box-Level Cues: Switching from mask IoU to bounding box IoU improves performance significantly, particularly in structured settings like MOT17. However, further gains from depth-segmentation integration indicate that relying on 2D overlap alone remains insufficient for robust identity tracking, especially under occlusion.

Role of Depth-Segmentation Association: Adding the encoder-refined depth-segmentation embeddings to appearance and bounding box cues yields consistent gains across both datasets. The improvement is especially pronounced on DanceTrack, where motion is more irregular and visual ambiguity is higher, highlighting their complementary value beyond 2D geometry.

3.5 Qualitative Results on Multiple Datasets

We present qualitative results of our SelfTrEncMOT tracker on three benchmarks: SportsMOT (See Figure 3.3), DanceTrack (See Figure 3.4), and MOT17 (See Figure 3.5). These datasets differ in motion complexity, occlusion, and crowd density. Our SelfTrEncMOT maintains consistent identity association across diverse scenarios, highlighting its robustness to motion variation, visual ambiguity, and occlusion.

Figure 3.3: Qualitative results of SelfTrEncMOT on the SportsMOT dataset. Despite rapid motion and visual clutter, SelfTrEncMOT yields strong identity association across time.

3.6 Summary

We presented SelfTrEncMOT, a novel tracking method that incorporates encoder-refined depth-segmentation embeddings as a standalone cue for robust object association. By fusing zero-shot monocular depth with promptable segmentation and refining the result via self-supervised learning, our method captures fine-grained spatial and geometric context often missed by traditional motion or appearance cues. Evaluations on diverse benchmarks confirm that SelfTrEncMOT is stable in identity preservation under occlusion, visual ambiguity, and crowd density. This work highlights

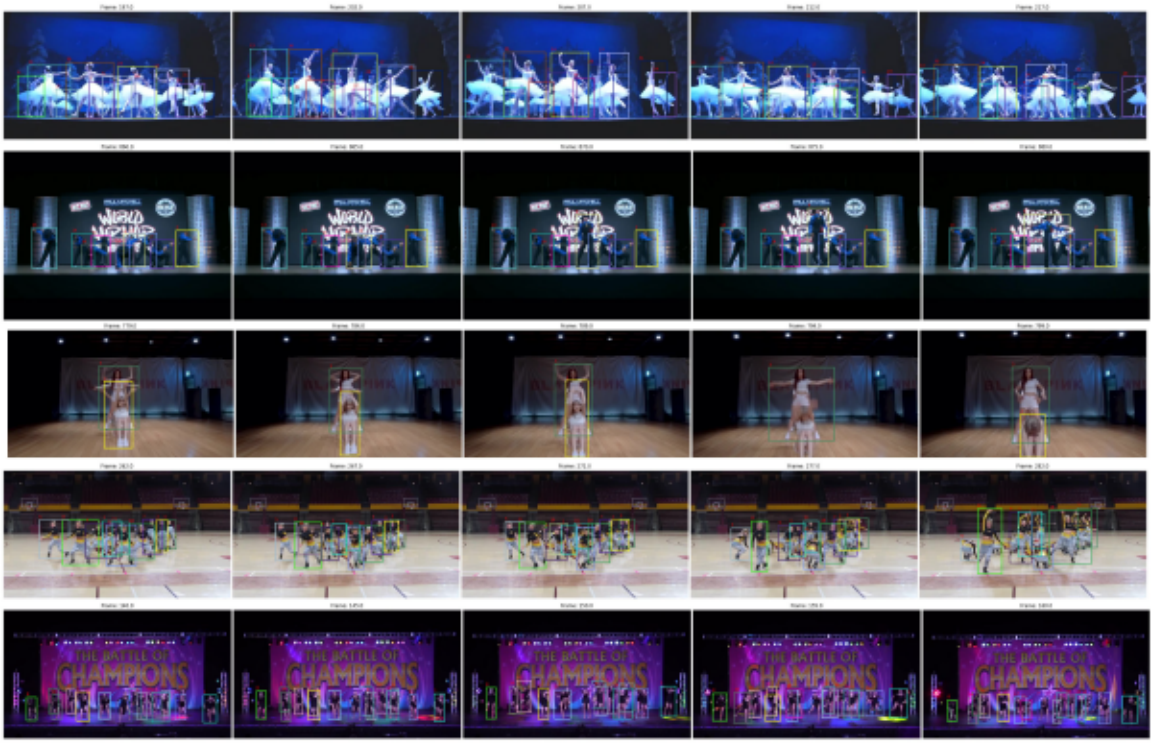

Figure 3.4: Tracking visualization on the DanceTrack dataset. The model demonstrates robust identity preservation in crowded and fast-motion scenarios.

Figure 3.5: Qualitative tracking results on MOT17. SelfTrEncMOT maintains consistent identities across occlusions.

the effectiveness of pixel-aligned, depth-aware reasoning in advancing multi-object tracking performance.

3.7 Comparison of Proposed Frameworks

Table 3.5 presents a comparative evaluation of our proposed Multi-Object Tracking frameworks *DepthMOT* and *SelfTrEncMOT*. While DepthMOT offers stronger association accuracy on DanceTrack, SelfTrEncMOT consistently achieves competitive or superior performance across all datasets with significantly higher runtime speed, which is mainly due to the avoidance of calculating IoU for segmentation masks.

Method	**DanceTrack**				**SportsMOT**			**MOT17**		
	HOTA ↑	AssA ↑	IDF1 ↑	FPS ↑	HOTA ↑	AssA ↑	IDF1 ↑	HOTA ↑	AssA ↑	IDF1 ↑
DepthMOT	**64.27**	**51.04**	**66.54**	3.35	76.2	65.6	76.9	62.7	63.6	77.9
SelfTrEncMOT	64.14	50.85	66.47	**125**	**76.4**	**66.0**	**77.1**	**63.48**	**63.25**	**78.12**

Table 3.5: Performance comparison between DepthMOT and SelfTrEncMOT on three benchmarks. Bold indicates best results.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

In this thesis, we introduced two novel tracking frameworks designed to address persistent challenges in Multi-Object Tracking (MOT), including non-linear motion, occlusion, and visual ambiguity. Both frameworks leverage geometric cues, specifically monocular depth and segmentation, to go beyond conventional 2D appearance-based association, offering more robust identity preservation in dynamic scenes.

The first framework, DepthMOT, integrates zero-shot monocular depth estimation into the association process and introduces a novel Hierarchical Alignment Score (HAS) that combines bounding box IoU with pixel-level alignment. This approach enables strong spatial differentiation between visually similar objects, especially in crowded or overlapping scenarios, without requiring any task-specific training or fine-tuning. The code is available at: https://github.com/Milad-Khanchi/DepthMOT.

The second framework, SelfTrEncMOT, focuses on enhancing association robustness through a self-supervised encoder that processes combined depth-segmentation features. By improving the temporal stability and discriminability of these spatial embeddings, this method provides a lightweight but effective representation for pixel-aligned matching, reducing the reliance on computationally expensive mask-IoU operations. The code can be found at: https://github.com/Milad-Khanchi/SelfTrEncMOT.

Comprehensive evaluations on benchmark datasets, DanceTrack and SportsMOT, demonstrate the effectiveness of both approaches. While DepthMOT achieves state-of-the-art or competitive results in association-based metrics such as HOTA, IDF1, and AssA, particularly in non-linear motion settings, SelfTrEncMOT offers competitive performance with improved stability under occlusion and appearance ambiguity. Together, these methods highlight the value of incorporating monocular depth and refined spatial features into the MOT pipeline, offering new directions for robust, geometry-aware tracking.

4.2 Future Work

While the proposed frameworks demonstrate competitive performance in both crowded and structured scenes, several avenues remain open for future exploration:

- **Temporal Modeling of Depth Features:** Current depth maps are used on a per-frame basis without explicit modeling of temporal consistency. Future work could explore recurrent or transformer-based architectures to exploit the sequential nature of depth evolution.
- **End-to-End Integration:** Our current pipeline relies on external modules for depth estimation and segmentation. A unified, end-to-end trainable model that jointly learns depth-aware embeddings and association logic could yield further improvements in speed and robustness.
- **Multi-Modal Fusion:** Incorporating other sensory cues such as optical flow, stereo depth, or LiDAR could enhance robustness in domains like autonomous driving and surveillance where multimodal data is available.
- **Domain Adaptation and Generalization:** Evaluating and extending the model to unseen environments, such as night-time scenes, remains an important step to ensure broader applicability.
- **Selective Tracking for Region-of-Interest (ROI):** Current MOT pipelines process all detected objects in a scene. However, in real-world applications such as sports analytics, surveillance, or driver monitoring, it is often sufficient to track specific objects within predefined

regions of interest (ROI). Future adaptations of our framework could incorporate spatial filtering or ROI-aware mechanisms to restrict association to relevant areas. This would reduce computational cost and enhance real-time performance.

In summary, this work establishes a promising direction in MOT by demonstrating how geometric context and pixel-aligned representations can be leveraged effectively for identity preservation, especially under visual complexity. Future advancements in this direction hold the potential to bridge the gap between appearance- and structure-based tracking paradigms.

Bibliography

[1] W. Lv, Y. Huang, N. Zhang, R.-S. Lin, M. Han, and D. Zeng, "DiffMot: A Real-time Diffusion-based Multiple Object Tracker with Non-linear Prediction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 19 321–19 330.

[2] G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani, "Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification," in *IEEE Int. Conf. Image Process.* IEEE, 2023, pp. 3025–3029.

[3] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9686–9696.

[4] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMot: On the Fairness of Detection and Re-identification in Multiple Object Tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 3069–3087, 2021.

[5] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus, and X. Alameda-Pineda, "TransCenter: Transformers with Dense Representations for Multiple-Object Tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7820–7835, 2022.

[6] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-Object Tracking with Transformers," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 8844–8854.

[7] G. Brasó, O. Cetintas, and L. Leal-Taixé, "Multi-Object Tracking and Segmentation Via Neural Message Passing," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 3035–3053, 2022.

[8] C. Du, C. Lin, R. Jin, B. Chai, Y. Yao, and S. Su, "Exploring the State-of-the-Art in Multi-Object Tracking: a Comprehensive Survey, Evaluation, Challenges, and Future Directions," *Multimedia tools and applications*, vol. 83, no. 29, pp. 73 151–73 189, 2024.

[9] J. Kiran and D. Ganesh, "A Comprehensive Study on Multiple Object Detection and Tracking in Videos Under Complex Environments Using Advanced Deep Learning Models," in *Int. Conf. Device Intell. Comput. Commun. Technol.* IEEE, 2025, pp. 107–112.

[10] G. Brasó and L. Leal-Taixé, "Learning a Neural Solver for Multiple Object Tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 6247–6257.

[11] O. Cetintas, G. Brasó, and L. Leal-Taixé, "Unifying Short and Long-Term Tracking with Graph Hierarchies," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 22 877–22 887.

[12] R. Gao and L. Wang, "MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking," in *Int. Conf. Comput. Vis.*, 2023, pp. 9901–9910.

[13] K. G. Quach, P. Nguyen, C. N. Duong, T. D. Bui, and K. Luu, "Depth Perspective-Aware Multiple Object Tracking," in *Eng. Appl. AI Swarm Intell.* Springer, 2024, pp. 181–205.

[14] J. Wang, H. Zheng, Y. Yu, Y. He, and Y. Liu, "Robust multiple obstacle tracking method based on depth aware OCSORT for agricultural robots," *Comput. Electron. Agric.*, vol. 217, p. 108580, 2024.

[15] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, and P. Luo, "DanceTrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 20 993–21 002.

[16] Y. Cui, C. Zeng, X. Zhao, Y. Yang, G. Wu, and L. Wang, "SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes," in *Int. Conf. Comput. Vis.*, 2023, pp. 9921–9931.

[17] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[18] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "MOT20: A benchmark for multi object tracking in crowded scenes," *arXiv preprint arXiv:2003.09003*, 2020.

[19] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "HOTA: A Higher Order Metric for Evaluating Multi-object Tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 548–578, 2021.

[20] L. Vaquero, Y. Xu, X. Alameda-Pineda, V. M. Brea, and M. Mucientes, "Lost and Found: Overcoming Detector Failures in Online Multi-object Tracking," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 448–466.

[21] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, and W. Hu, "Rethinking the Competition Between Detection and ReID in Multiobject Tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 3182–3196, 2022.

[22] M. Khanchi, M. Amer, and C. Poullis, "Depth-aware scoring and hierarchical alignment for multiple object tracking," in *IEEE Int. Conf. Image Process.*, 2025, accepted.

[23] K. Shim, J. Hwang, K. Ko, and C. Kim, "A Confidence-Aware Matching Strategy for Generalized Multi-Object Tracking," in *IEEE Int. Conf. Image Process.* IEEE, 2024, pp. 4042–4048.

[24] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online Multi-Object Tracking with Dual Matching Attention Networks," in *Eur. Conf. Comput. Vis.*, 2018, pp. 366–382.

[25] Z. Lu, B. Shuai, Y. Chen, Z. Xu, and D. Modolo, "Self-Supervised Multi-Object Tracking with Path Consistency," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[26] Y. Zhang, T. Wang, and X. Zhang, "MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 22 056–22 065.

[27] Z. Ge, "YOLOX: Exceeding YOLO Series in 2021," *arXiv:2107.08430*, 2021.

[28] C. Huang, S. Han, M. He, W. Zheng, and Y. Wei, "DeconfuseTrack: Dealing with Confusion for Multi-Object Tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 19 290–19 299.

[29] C.-J. Liu and T.-N. Lin, "DET: Depth-Enhanced Tracker to Mitigate Severe Occlusion and Homogeneous Appearance Problems for Indoor Multiple-Object Tracking," *IEEE Access*, vol. 10, pp. 8287–8304, 2022.

[30] G. Zheng, L. Songtao, W. Feng, L. Zeming, S. Jian *et al.*, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[31] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object Tracking by Associating Every Detection Box," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 1–21.

[32] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A Pytorch Toolbox for General Instance Re-identification," in *ACM Int. Conf. Multimedia*, 2023, pp. 9664–9667.

[33] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," *arXiv preprint arXiv:2206.14651*, 2022.

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

[35] T. Delatolas, V. Kalogeiton, and D. P. Papadopoulos, "Learning the What and How of Annotation in Video Object Segmentation," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2024, pp. 6951–6961.

[36] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 5559–5568.

[37] Y. Heo, Y. Jun Koh, and C.-S. Kim, "Interactive Video Object Segmentation Using Global and Local Transfer Modules," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 297–313.

[38] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "SAM 2: Segment Anything in Images and Videos," *arXiv preprint arXiv:2408.00714*, 2024.

[39] V. Guizilini, I. Vasiljevic, D. Chen, R. Ambruș, and A. Gaidon, "Towards Zero-Shot Scale-Aware Monocular Depth Estimation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9233–9243.

[40] S. Saxena, J. Hur, C. Herrmann, D. Sun, and D. J. Fleet, "Zero-Shot Metric Depth with a Field-of-View Conditioned Diffusion Model," *arXiv preprint arXiv:2312.13252*, 2023.

[41] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth Pro: Sharp Monocular Metric Depth in Less Than a Second," *arXiv preprint arXiv:2410.02073*, 2024.

[42] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *IEEE Int. Conf. Image Process.* IEEE, 2016, pp. 3464–3468.

[43] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 17–35.

[44] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.

[45] R. Luo, Z. Song, L. Ma, J. Wei, W. Yang, and M. Yang, "DiffusionTrack: Diffusion Model for Multi-Object Tracking," in *AAAI*, vol. 38, no. 5, 2024, pp. 3991–3999.

[46] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-End Multiple-Object Tracking with Transformer," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 659–675.

[47] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking Objects as Points," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 474–490.

[48] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan, "Track To Detect and Segment: An Online Multi-Object Tracker," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 12 352–12 361.

[49] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: Multiple Object Tracking with Transformer," *arXiv preprint arXiv:2012.15460*, 2020.

[50] J. Seidenschwarz, G. Brasó, V. C. Serrano, I. Elezi, and L. Leal-Taixé, "Simple Cues Lead to a Strong Multi-Object Tracker," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 13 813–13 823.

[51] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE Int. Conf. Image Process.* IEEE, 2017, pp. 3645–3649.

[52] C. Xiao, Q. Cao, Y. Zhong, L. Lan, X. Zhang, Z. Luo, and D. Tao, "MotionTrack: Learning motion predictor for multiple object tracking," *Neural Networks*, vol. 179, p. 106539, 2024.

[53] Z. Qin, L. Wang, S. Zhou, P. Fu, G. Hua, and W. Tang, "Towards Generalizable Multi-Object Tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 18 995–19 004.

[54] F. Yang, S. Odashima, S. Masui, and S. Jiang, "Hard to Track Objects With Irregular Motions and Similar Appearances? Make It Easier by Buffering the Matching Space," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4799–4808.

[55] X. Zhou, T. Yin, V. Koltun, and P. Krähenbühl, "Global Tracking Transformers," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 8771–8780.

[56] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu, "Quasi-Dense Similarity Learning for Multiple Object Tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 164–173.

[57] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4870–4880.

[58] Y. Wang, K. Kitani, and X. Weng, "Joint Object Detection and Multi-Object Tracking with Graph Neural Networks," in *IEEE Int. Conf. Robot. Autom.* IEEE, 2021, pp. 13 708–13 715.

[59] S. Li, L. Yang, H. Tan, B. Wang, W. Huang, H. Liu, W. Yang, and L. Lan, "Self-supervised re-identification for online joint multi-object tracking," *Knowledge and Information Systems*, vol. 67, no. 1, pp. 881–914, 2025.

[60] D. C. Bui, H. A. Hoang, and M. Yoo, "AFMtrack: Attention-Based Feature Matching for Multiple Object Tracking," *IEEE Access*, 2024.

[61] F. Wu and Y. Zhang, "DilateTracker: Feature Fusion Dilated Transformer for Multi-object Tracking," in *2024 4th International Symposium on AI (ISAI)*. IEEE, 2024, pp. 48–54.