applied sciences

MDPI

Article

KalmanFormer: Integrating a Deep Motion Model into SORT for Video Multi-Object Tracking

Jiayu Hong [1], Yunyao Li [1], Jielu Yan [1,*], Xuekai Wei [1,*], Weizhi Xian [1,*] and Yi Qin [2]

[1] College of Computer Science, Chongqing University, Chongqing 400044, China
[2] College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China
* Correspondence: yanjielu@cqu.edu.cn (J.Y.); xuekaiwei2-c@my.cityu.edu.hk (X.W.); wasxxwz@163.com (W.X.)

Abstract

This paper presents the study of integrating a deep motion model into simple online and real-time tracking for video multi-object tracking. The tracking-by-detection paradigm faces significant challenges in handling nonlinear motion and occlusions. Although conventional Kalman-filter-based methods such as the SORT are efficient, they suffer from error accumulation because of their linear motion assumption. We propose KalmanFormer, a novel framework that enhances Kalman-filter-based tracking through adaptive motion modeling for video sequences. KalmanFormer consists of three key components. First, the inner-trajectory motion corrector, built upon the transformer architecture, refines Kalman filter predictions by learning nonlinear residuals from historical trajectories, thereby improving adaptability to complex motion patterns in videos. Second, the cross-trajectory attention module captures interobject motion correlations, significantly boosting object association under occlusions. Third, a pseudo-observation generator is integrated to provide neural-based predictions when detections are missing, stabilizing the Kalman filter update process. To validate our approach, we conduct comprehensive evaluations on the video benchmarks DanceTrack, MOT17, and MOT20 to demonstrate its effectiveness in handling complex motion and occlusion. The experimental results on the DanceTrack, MOT17, and MOT20 benchmarks demonstrate that KalmanFormer achieves competitive performance, with HOTA scores of 66.6 on MOT17 and 63.2 on MOT20, and strong identity preservation, IDF1: 82.0% and 80.1%, respectively.

Keywords: video multi-object tracking; Kalman filter; transformer; nonlinear motion; occlusion handling; adaptive motion modeling

check for updates

Academic Editor: Andrea Prati

Received: 10 August 2025
Revised: 1 September 2025
Accepted: 2 September 2025
Published: 4 September 2025

Citation: Hong, J.; Li, Y.; Yan, J.; Wei, X.; Xian, W.; Qin, Y. KalmanFormer: Integrating a Deep Motion Model into SORT for Video Multi-Object Tracking. *Appl. Sci.* **2025**, *15*, 9727. https://doi.org/10.3390/app15179727

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Video multi-object tracking (video MOT) is fundamental to dynamic scene understanding, supporting applications such as autonomous navigation, video surveillance, and crowd analysis. Owing to their simplicity and efficiency, the dominant tracking-by-detection paradigm frequently employs Kalman-filter-based models [1–7]. However, these methods assume linear motion, making them ill suited for real-world scenarios characterized by nonlinear trajectories [8] and frequent occlusions [9]. As a result, issues such as identity switches, track fragmentation, and target loss become prominent, particularly in crowded scenes [10]. Although many improvements have been proposed, state-of-the-art KF-based trackers such as simple online and real-time tracking (SORT) [1] and OC-SORT [9] remain limited by their constant-velocity motion assumption [3]. This makes them inadequate for handling abrupt trajectory changes common in applications such as

sports analytics [11] and pedestrian tracking [12]. Nonlinear variants, such as the extended Kalman filter [13] and the unscented Kalman filter [14], attempt to address this gap but are constrained by Gaussian noise assumptions and increased computational cost [15], limiting their real-time applicability [16–19].

Another challenge arises from the short-term nature of traditional KF updates, which rely solely on the latest observations and neglect long-term motion dependencies. This limitation is particularly detrimental under occlusion, where the absence of historical context degrades tracking robustness [20]. While transformer-based models have succeeded in sequence modeling [21], their high computational overhead hinders real-time deployment [22]. Moreover, KF-based trackers suffer from error accumulation during prolonged occlusions, as predictions proceed without external corrections [9]. Although particle filters offer a more flexible framework, their computational complexity makes them impractical for large-scale tracking [23,24].

To address the aforementioned limitations, we propose KalmanFormer, a hybrid multi-object tracking (MOT) framework that integrates deep motion modeling into a traditional Kalman-filter-based pipeline. Unlike existing approaches that rely solely on linear motion assumptions, KalmanFormer introduces transformer-based components to enhance motion estimation and association under complex dynamics and occlusion. Specifically, the inner-trajectory motion corrector (ITMC) leverages a transformer pretrained on individual trajectories to learn nonlinear motion residuals, mitigating the limitations of the KF's linear prediction. The cross-trajectory attention module (CTAM) captures interactions across object trajectories, improving the association performance during occlusions. Furthermore, a pseudo-observation generator (POG) synthesizes surrogate observations when detections are missing, preventing drift and error accumulation. In contrast to prior transformer-based MOT methods that process raw image features [25–28], KalmanFormer operates directly on sequences of bounding boxes via lightweight attention mechanisms. Extensive experiments on the MOT17, MOT20, and DanceTrack benchmarks demonstrate that KalmanFormer achieves state-of-the-art performance while maintaining real-time efficiency, confirming its effectiveness in complex environments.

Table 1 provides a comparative overview of existing tracking methods, highlighting their key characteristics and limitations compared with our proposed KalmanFormer approach.

Table 1. Comparison of different tracking approaches. The table summarizes the key characteristics and limitations of existing tracking methods compared with our proposed KalmanFormer approach.

Method Type	Representative Works	Key Characteristics & Limitations
Traditional TBD (Tracking-by-Detection)	SORT [1], DeepSORT [2], ByteTrack [29], OC-SORT [9]	Efficient computation, real-time performance; poor handling of nonlinear motion, identity switches during occlusions
End-to-End Transformer (Image-to-Track)	TransTrack [25], TrackFormer [21], MOTR [30], MeMOTR [31]	Strong feature representation, joint detection and tracking; high computational overhead, slow inference speed
KalmanFormer (Hybrid TBD + Transformer)	Ours	Combines KF efficiency with transformer's nonlinear modeling, handles occlusions via pseudo-observations; relies on geometric information

Our main contributions are summarized as follows:

- We propose KalmanFormer, a hybrid tracking framework that integrates transformer-based modules with the Kalman filter. It incorporates an inner-trajectory motion

corrector (ITMC) and a cross-trajectory attention module (CTAM) for modeling nonlinear motion and interobject interactions.
- We introduce a pseudo-observation generator that addresses detection failures by producing reliable surrogate observations, enhancing robustness under occlusions and missing detections.
- We conduct comprehensive experiments on the MOT17, MOT20, and DanceTrack benchmarks. Compared with existing SORT variants, the KalmanFormer achieves state-of-the-art performance and significantly reduces the number of identity switches, particularly in occlusion-heavy scenarios.

2. Related Work

2.1. Single-Object Tracking

Single-object tracking (SOT) involves tracking a target across a video, facing challenges such as occlusion, scale changes, and background clutter. Early methods, such as correlation filters [16,17,23,32,33], achieved success in terms of computational efficiency. Some approaches have extended this to multimodal tracking by combining RGB and thermal data, which has improved performance under low-light conditions [24,27,28,34–37]. Particle filter methods [19,38–40] were introduced to address noisy environments, but they can be slow and not suitable for real-time tracking [18,29,41,42].

With the rise of deep learning, trackers such as GOTURN [43] and SiamFC [44] improved speed and accuracy by using convolutional neural networks (CNNs) for feature extraction [4–6,45–47]. Some newer methods, such as SCOOT [48], use self-supervised learning and the fusion of multiple data sources to handle more complex situations [7,49–51]. Despite their progress, appearance-based trackers are still vulnerable to occlusions and large changes in scale, pose, or lighting. To address this, methods such as ATOM [52] add motion models for better robustness [53–55], whereas SiamRPN++ [56] helps with scale changes by combining region proposal networks [57–59].

More recently, transformer-based models, such as TransT [60], have been used for SOT [33,39], but they often face the challenge of high computational cost, limiting their use in real-time applications [61–65]. These methods could be improved through model compression techniques, sparse attention mechanisms, or hardware-specific optimizations to reduce their computational overhead while maintaining their strong representational capabilities [66–69].

2.2. Multi-Object Tracking

Multi-object tracking (MOT) extends SOT by tracking multiple targets simultaneously, which adds complexity, especially in dense or occlusion-heavy scenes. Some MOT trackers, such as CenterTrack [70], focus only on motion and predict positions on the basis of minimal information but struggle to recover occluded targets [71–73]. Other methods focus on maintaining object identity during occlusions but rely on learned heuristics [33].

SORT-based methods [1,2,9,29,41,42,45] use Kalman filters for motion prediction and the Hungarian algorithm for data association. These methods are efficient but fail in situations where motion is nonlinear or when objects are occluded for long periods. DeepSORT [2] and ByteTrack [29] improve performance by adding appearance cues but still suffer from the linear motion assumption of Kalman filters. Transformer-based MOT methods, such as TransTrack [25,49–51], TrackFormer [21,53–55], and MOTR [30,57–59], attempt to overcome these problems by using attention mechanisms to model motion and temporal dependencies more effectively [37,61–63].

Recent advances in deep metric learning and cross-modal tracking [64–67] have also contributed to improved performance in challenging scenarios [68,69,71–73].

3. Method

We propose KalmanFormer, a hybrid multi-object tracking framework that integrates the strengths of Kalman filters and transformers for robust motion modeling and occlusion handling. As shown in Figure 1, the framework consists of three key modules: the inner-trajectory motion corrector (ITMC), which captures object-specific nonlinear motion patterns from historical trajectories; the cross-trajectory attention module (CTAM), which models interactions between nearby objects to refine predictions in crowded environments; and the pseudo-observation generator (POG), which provides surrogate measurements when detections are missing. These modules are integrated into a unified Kalman-filter-based framework, where ITMC and CTAM refine motion estimates, and the pseudo-observation mechanism ensures continuity during occlusions.

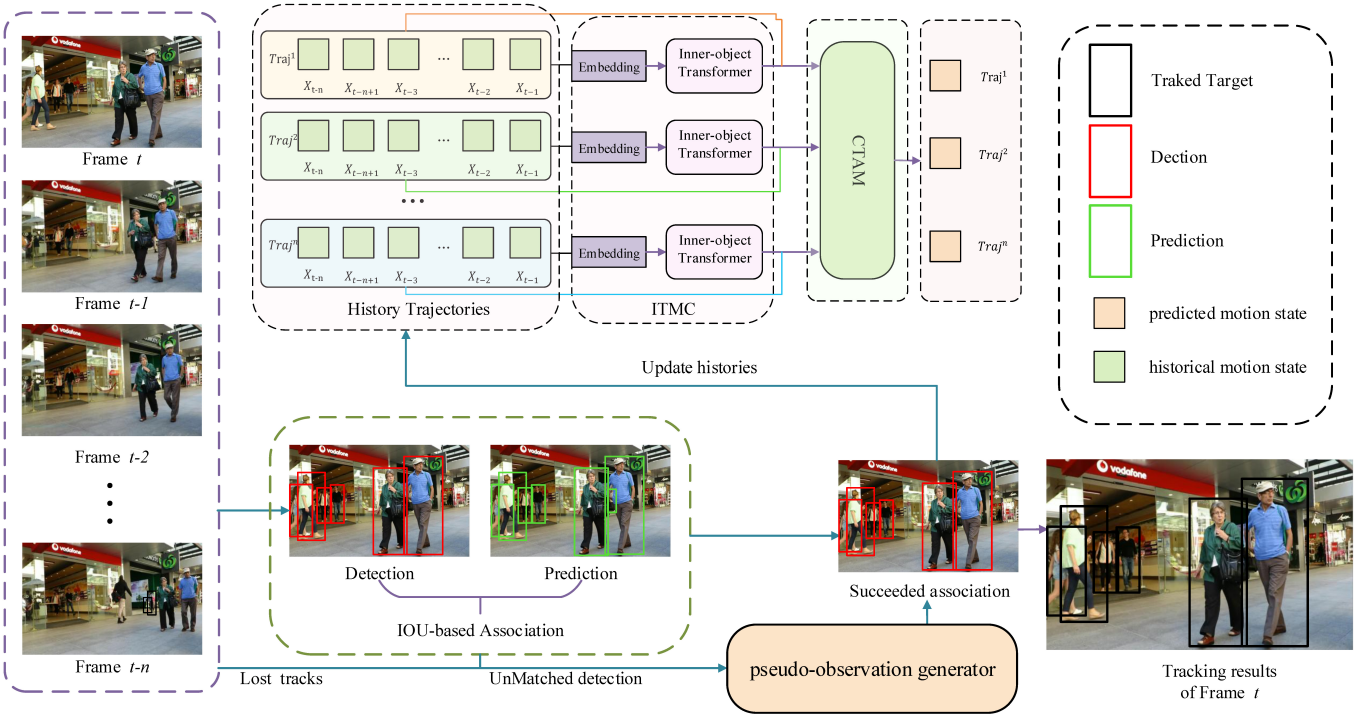


Figure 1. Overview of the KalmanFormer framework. The checkered rectangle on the right represents the figure caption, illustrating the key components and data flow of our approach.

3.1. Notation and State Definition

We adopt an absolute bounding-box state representation for each object at time t as $\mathbf{x}_t = [c_x, c_y, a, r, v_x, v_y, v_a]^\top$, where (c_x, c_y) denotes box center coordinates in pixels, a denotes the box area, r is the aspect ratio, and (v_x, v_y, v_a) represents velocities. The standard constant-velocity Kalman model is used with process noise $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, Q)$ and measurement noise $\boldsymbol{\eta}_t \sim \mathcal{N}(0, R)$. Throughout, $\mathbf{x}_t^{\text{inner}}$ denotes the ITMC prediction (absolute coordinates), $\hat{\mathbf{x}}_t^{\text{hybrid}}$ denotes the hybrid motion prediction before Kalman correction, and $\mathbf{x}_t^{\text{final}}$ denotes the posterior state after the Kalman update. Unless otherwise stated, all coordinates are absolute in the image frame. When we refer to residuals, these are additive residuals in the absolute state space. The same MLP used in Equation (1) is reused after CTAM to map interaction-aware embeddings to the absolute state space.

3.2. Hybrid Motion Prediction and Transformer Integration

Kalman filters assume linear dynamics, which often fail in real-world tracking scenarios involving acceleration or abrupt turns. To overcome this, we introduce a hybrid

motion prediction model that augments the Kalman filter's output with a nonlinear residual learned by a transformer:

$$\hat{\mathbf{x}}_t^{\text{hybrid}} = \underbrace{\mathbf{F}\mathbf{x}_{t-1}}_{\text{Linear prediction}} + \underbrace{\text{Transformer}(\mathbf{x}_{t-N:t-1})}_{\text{Nonlinear residual}}, \tag{1}$$

where the residual is applied in absolute state space; process noise is handled by the Kalman filter via $\boldsymbol{\varepsilon}_t$.

The core innovation of KalmanFormer lies in the deep integration of a transformer architecture within the recursive Bayesian estimation framework of the Kalman filter. This hybrid approach is designed to leverage the transformer's superior ability to model complex, nonlinear dynamics and long-range dependencies from data while retaining the Kalman filter's efficiency and probabilistic rigor for state estimation. The integration is realized through three specialized components that augment the classical Kalman filter cycle: a trajectory corrector, an interobject attention module, and a pseudo-observation generator.

3.2.1. Mathematical Formulation of Integration

The standard Kalman filter consists of a prediction step and an update step. KalmanFormer enhances both steps by integrating transformer-based components that learn complex motion patterns and interobject interactions from data. The integration is achieved through two key components:

$$\mathbf{x}_t^{\text{corrected}} = \underbrace{\mathbf{F}\mathbf{x}_{t-1}}_{\text{KF prediction}} + \underbrace{\text{ITMC}(\mathbf{x}_{t-N:t-1})}_{\text{Nonlinear correction}} \tag{2}$$

where ITMC processes the trajectory history to predict nonlinear motion patterns and:

$$\mathbf{K}_t = \mathbf{K}_t^{\text{base}} \cdot \text{CTAM}(\{\mathbf{x}_t^i\}_{i=1}^M) \tag{3}$$

where CTAM models interobject interactions to adaptively adjust the Kalman gain. The specific architectures and implementations of ITMC and CTAM are detailed in their respective sections.

3.2.2. Limitations of the Integrated Model

Despite its strong performance, the KalmanFormer architecture has several inherent limitations. The interobject attention module has a computational complexity of $O(N^2)$, where N is the number of objects being tracked, which can become a bottleneck in scenarios with a very large number of objects, potentially hindering real-time performance. As a data-driven model, the performance of the transformer components is heavily dependent on the diversity and representativeness of the training data, and the model may struggle to generalize to motion patterns or object interaction behaviors that are drastically different from those seen during training. While the model is designed for nonlinear movements, its predictive capabilities may degrade when tracking objects with extremely erratic, aperiodic, or random motion profiles that lack any learnable underlying pattern, where the learned priors of the transformer might introduce bias. Additionally, the sequential processing of the transformer modules within the Kalman filter loop introduces additional latency compared with a standard Kalman filter, which must be considered for applications with stringent real-time constraints.

3.3. Inner-Trajectory Motion Corrector (ITMC)

The ITMC explicitly models long-term, nonlinear dynamics via a transformer encoder. As detailed in Figure 2, it takes the past N states of an object as input and encodes them into a latent representation:

$$\mathbf{Z}_t = \text{TransformerEncoder}(\text{Embed}(\mathbf{x}_{t-N:t-1})), \tag{4}$$

where $\text{Embed}(\cdot)$ maps each input state to a high-dimensional vector and $\mathbf{Z}_t \in \mathbb{R}^{N \times d'}$ denotes the encoded sequence. We extract the last token (or a [CLS]-style representation) $\mathbf{Z}_t^{[\text{last}]} \in \mathbb{R}^{d'}$ and produce a prediction via a multilayer perceptron (MLP):

$$\hat{\mathbf{x}}_t^{\text{inner}} = \text{MLP}(\mathbf{Z}_t^{[\text{last}]}). \tag{5}$$

This nonlinear prediction is fused with the Kalman linear output in a hybrid estimate, improving robustness under curved, abrupt, and stop-and-go motions.

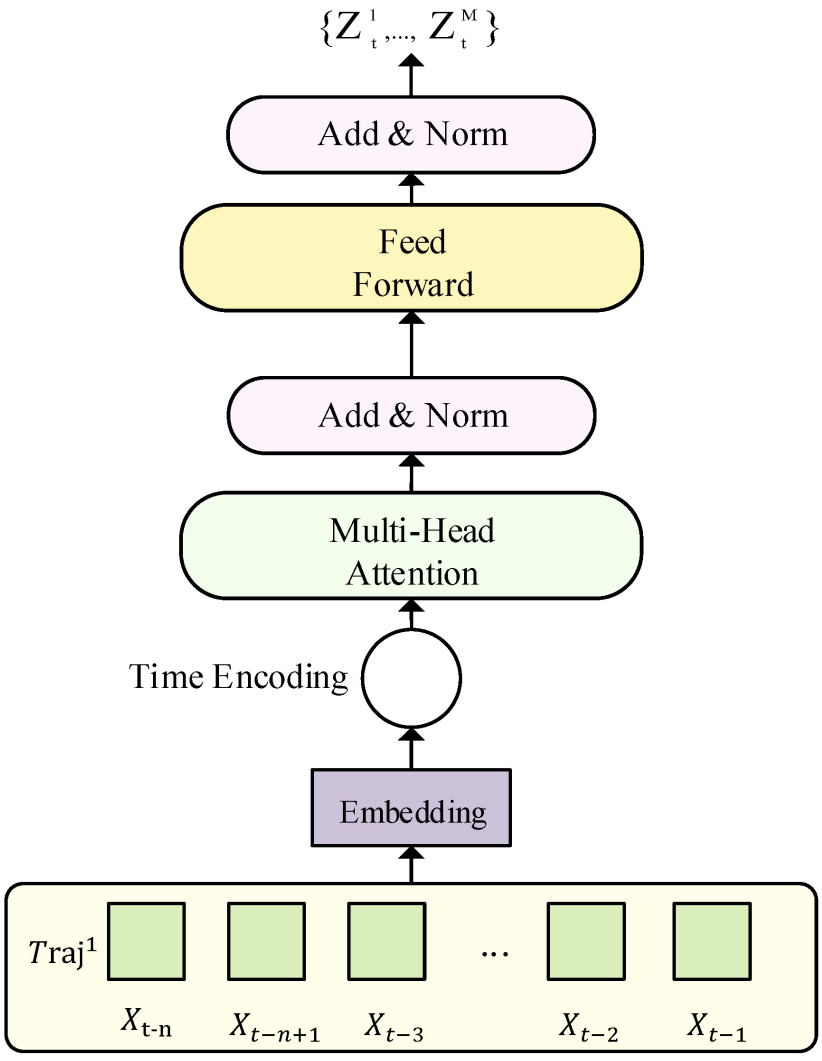


Figure 2. The architecture of the inner-trajectory motion corrector (ITMC).

3.4. Cross-Trajectory Attention Module (CTAM)

CTAM models interactions among multiple targets via attention, as illustrated in Figure 3, yielding interaction-aware embeddings for refined prediction and association. Given that M targets at time t with embeddings $\{\mathbf{Z}_t^1, \ldots, \mathbf{Z}_t^M\}$, each $\mathbf{Z}_t^i$ is projected to

queries ($\mathbf{Q}_i$), keys ($\mathbf{K}_j$), and values ($\mathbf{V}_j$) via learned weight matrices $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$, as shown in Equation (6):

$$\mathbf{Q}_i = \mathbf{W}_q \mathbf{Z}_t^i, \quad \mathbf{K}_j = \mathbf{W}_k \mathbf{Z}_t^j, \quad \mathbf{V}_j = \mathbf{W}_v \mathbf{Z}_t^j. \tag{6}$$

The attention score α_{ij} between object i and object j is defined as the scaled dot product, as shown in Equation (7):

$$\alpha_{ij} = \frac{\mathbf{Q}_i \cdot \mathbf{K}_j^\top}{\sqrt{d_a}}, \tag{7}$$

where d_a is the dimension of the attention space. We adopt a top-K sparse attention mechanism to retain only the most relevant K neighbors for each target, as shown in Equation (8):

$$\mathcal{N}_K(i) = \mathrm{TopK}\left(\{\alpha_{ij}\}_{j=1}^{M}\right), \quad \tilde{\alpha}_{ij} = \begin{cases} \alpha_{ij}, & j \in \mathcal{N}_K(i) \\ -\infty, & \text{otherwise} \end{cases} \tag{8}$$

The interaction-enhanced embedding is computed as a weighted sum of the values, as shown in Equation (9):

$$\hat{\mathbf{Z}}_t^i = \sum_{j \in \mathcal{N}_K(i)} \mathrm{Softmax}(\tilde{\alpha}_{ij}) \cdot \mathbf{V}_j, \tag{9}$$

where Softmax normalizes the attention scores to create a probability distribution over the selected neighbors. This embedding is then mapped back to the absolute state space for hybrid prediction and association.

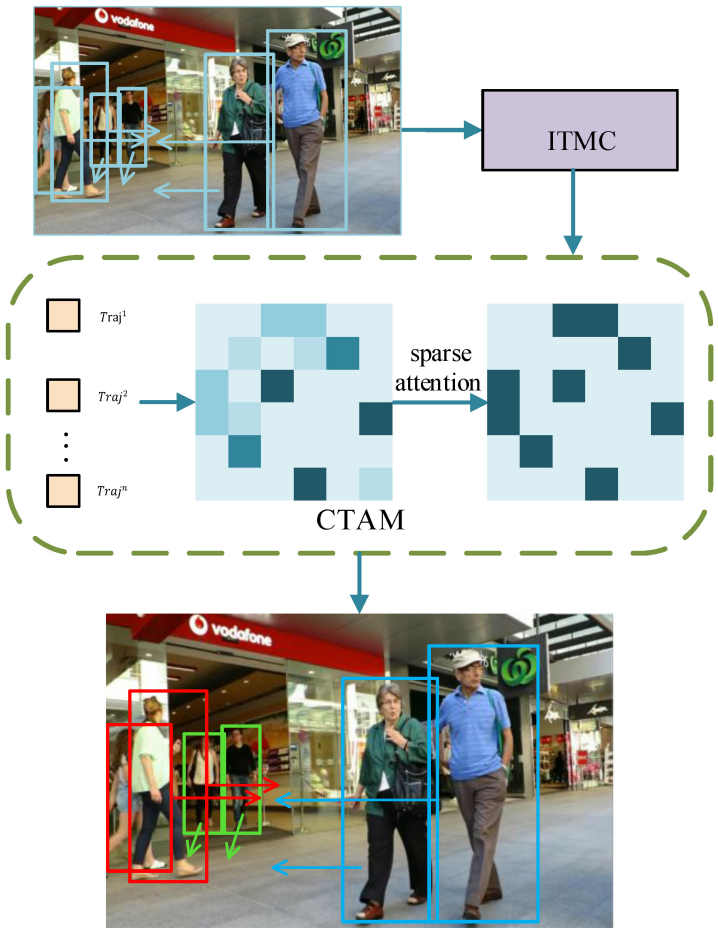


Figure 3. Illustration of the cross-trajectory attention module (CTAM).

3.5. Pseudo-Observation Generation

When detections are available, the hybrid prediction serves as the motion prior to the Kalman prediction step. When detections are missing or unreliable, we substitute a pseudo-observation to avoid error accumulation under occlusion, as defined in Equation (10):

$$\mathbf{o}_t = \hat{\mathbf{x}}_t^{\text{hybrid}} \tag{10}$$

where $\mathbf{o}_t$ is the generated pseudo-observation at time t. The Kalman update then proceeds as usual, as shown in Equation (11):

$$\mathbf{x}_t = \mathbf{x}_{t|t-1} + \mathbf{K}_t\left(\mathbf{o}_t - \mathbf{H}\,\mathbf{x}_{t|t-1}\right) \tag{11}$$

where $\mathbf{K}_t$ is the Kalman gain at time t and where $\mathbf{H}$ is the observation matrix. The key insight is that our pseudo-observation is not merely a simple linear state extrapolation. Instead, it is the direct output of our full hybrid motion model, as defined in Equation (1). This generated observation is therefore informed by both the object's own learned nonlinear dynamics via the ITMC and the contextual influence of its neighbors via the CTAM. This makes it a significantly more robust estimate than what a standard Kalman filter could produce on its own. In terms of accuracy under strong occlusions, the performance of the POG degrades gradually. For short-term occlusions (1–5 frames), the recent trajectory provides a strong prior, and the generated pseudo-observation is highly accurate. For longer-term occlusions, while the uncertainty of the prediction naturally increases over time, the learned deep motion model provides a much more physically plausible trajectory than does a constant-velocity assumption. This prevents catastrophic drift and increases the likelihood of correct reidentification when the object's detection becomes available again. The training targets for ITMC/CTAM are absolute states derived from the annotated boxes; the residual learner minimizes the discrepancy between $\hat{\mathbf{x}}_t^{\text{hybrid}}$ and the ground-truth absolute state. This keeps the notation consistent: $\mathbf{x}_t^{\text{inner}}$ and $\hat{\mathbf{x}}_t^{\text{hybrid}}$ are in the same coordinate space, and $\mathbf{x}_t^{\text{final}}$ is the KF posterior. We refer to this component as the pseudo-observation generator (POG).

3.6. Complete Tracking Pipeline

To synthesize these components into a complete tracking pipeline, we present Algorithm 1, which integrates the Kalman filter with our transformer-based components. The algorithm begins by predicting the next state for all existing tracklets via the standard Kalman filter prediction step. The ITMC module then refines these predictions by incorporating nonlinear motion patterns learned from historical trajectories, followed by the CTAM module, which further enhances the predictions by modeling interobject interactions. The resulting motion-aware states are matched with incoming detections via the Hungarian algorithm. The matched pairs are updated directly with the detections, whereas the unmatched tracklets are updated via our pseudo-observation mechanism to maintain track continuity during occlusions. Finally, unmatched detections are used to initialize new tracklets. This integrated approach ensures robust tracking even in challenging scenarios with nonlinear motion and frequent occlusions.

Algorithm 1 KalmanFormer Tracking Pipeline

Input: Detections $\mathcal{D}_t$, existing tracklets $\mathcal{T}_{t-1}$
Output: Updated tracklets $\mathcal{T}_t$
for each $\tau_i \in \mathcal{T}_{t-1}$ **do**
  Predict $\hat{\tau}_i^t$ via Kalman filter
  Generate pseudo-observation $\tilde{o}_i^t$ if needed
end for
Fuse $\hat{\tau}^t$ and $\mathcal{D}_t$ to form candidate set $\mathcal{X}^t$
$\mathcal{X}' \leftarrow$ Inner-trajectory correction via ITMC
$\mathcal{X}'' \leftarrow$ Cross-trajectory refinement via CTAM
Compute the similarity matrix S between $\mathcal{X}''$ and $\mathcal{D}_t$
Apply matching (e.g., Hungarian) on S
for each matched pair (τ_i, d_j) **do**
  Update τ_i with d_j
end for
for each unmatched τ_i **do**
  if it is missing for N frames **then**
    Discard
  else
    Update with pseudo-observation
  end if
end for
for each unmatched d_j **do**
  Initialize a new tracklet
end for
Return: $\mathcal{T}_t$

3.7. Training Strategy

We train the inner-trajectory and intertrajectory modules via trajectory ground truths from public MOT datasets. The training process consists of two stages:

- Stage 1 (Pretraining): The inner trajectory module is trained to minimize the L2 distance between the hybrid prediction $\hat{\mathbf{x}}_t^{\text{hybrid}}$ and the ground truth $\mathbf{x}_t$ via single-object tracks.
- Stage 2 (Joint training): The intertrajectory attention module is introduced, and the model is fine-tuned on multi-object sequences. The loss includes both trajectory prediction loss and consistency loss for correct interactions.

The total loss consists of:

- Inner trajectory loss:

$$\mathcal{L}_{\text{inner}} = \left\| \text{MLP}(\mathbf{x}_t^{\text{inner}}) - \mathbf{x}_t \right\|_2^2 \tag{12}$$

- Intertrajectory loss:

$$\mathcal{L}_{\text{inter}} = \left\| \text{CrossAttn}(\mathbf{x}_t^{\text{inner}}) - \mathbf{x}_t \right\|_2^2 \tag{13}$$

- Hybrid output loss:

$$\mathcal{L}_{\text{hybrid}} = \left\| \mathbf{x}_t^{\text{final}} - \mathbf{x}_t \right\|_2^2 \tag{14}$$

Additionally, we simulate realistic tracking challenges during training: random frame drop, target switching, and target loss simulation by masking.

3.8. Comparison with Transformer-Based Methods

Several recent MOT methods, such as TransTrack, TrackFormer, and MOTR, have successfully employed transformers. It is important to distinguish KalmanFormer from these approaches.

The primary difference lies in the input domain. Methods such as TransTrack and MOTR are end-to-end models that operate directly on image features extracted by a CNN backbone. While powerful, this makes them computationally expensive and tightly couples the tracking logic to a specific visual detector. In contrast, KalmanFormer operates purely on geometric and kinematic data (i.e., sequences of bounding box coordinates). This is a deliberate design choice that decouples our motion model from the detection process and makes our framework significantly more lightweight and efficient, allowing for easier integration with any detector and better real-time performance. Our strong results on multiple benchmarks demonstrate that a sophisticated motion model can rival or even surpass feature-based methods, especially in motion-centric challenges.

3.9. Computational Complexity and Efficiency

KalmanFormer operates purely on geometry without image feature extraction. ITMC and CTAM adopt sparse top-K attention, yielding $\mathcal{O}(K \cdot M)$ complexity per frame for M tracked objects. This design is suitable for real-time or near real-time settings on modern GPUs/CPUs.

To assess the scalability of KalmanFormer, we analyze its theoretical computational complexity with respect to the number of tracked objects, M. The model's total complexity per frame is the sum of its main components: the Kalman filter updates, the Inner-Trajectory Motion Corrector (ITMC), and the Cross-Trajectory Attention Module (CTAM).

- Kalman Filter: The prediction and update steps for M objects have a complexity of $O(M \cdot d^2)$, where d is the dimension of the state vector. The calculations for each object are independent, making this component highly parallelizable.
- ITMC: The ITMC processes a fixed-length history of N states for each of the M objects. While the total computational load is linear in M, with a complexity of $O(M \cdot N^2 \cdot d')$ where d' is the embedding dimension, the key architectural advantage is its parallel nature. The processing of each trajectory is independent of the others. On modern hardware with sufficient parallel processing units (e.g., GPUs), the wall-clock time required for this step is largely determined by the computation for a single trajectory, i.e., approximately $O(N^2 \cdot d')$, and is therefore nearly constant for a wide range of M.
- CTAM: This module is the most critical for scalability. A naive, full self-attention mechanism would incur a prohibitive $O(M^2 \cdot d')$ complexity, rendering the model impractical for scenes with many objects. However, as described in our method, we employ a top-K sparse attention mechanism. Each object attends only to its K most relevant neighbors, where K is a small constant ($K \ll M$). This crucial design choice reduces the complexity of CTAM to $O(M \cdot K \cdot d')$, re-establishing a linear and manageable relationship with the number of objects. This operation is also efficiently implemented via parallel batch matrix multiplication on GPUs.

Consequently, the overall theoretical complexity of KalmanFormer is approximately $O(M \cdot (d^2 + N^2 \cdot d' + K \cdot d'))$. The model scales linearly with the number of tracked objects M, and its parallel-friendly design ensures that its practical runtime remains efficient, avoiding the quadratic bottleneck of standard transformer architectures. This makes it a highly scalable solution for real-world MOT tasks.

As established in our complexity analysis, KalmanFormer is designed for efficiency. By operating on low-dimensional geometric data instead of high-dimensional image features, its computational overhead is minimal. The model's parameters are fixed after offline training, meaning that there is no online adaptation or fine-tuning required during inference. This results in consistent and predictable latency, making it well suited for real-time applications where a stable frame rate is necessary. Our use of sparse attention in the CTAM is critical to ensuring this real-time capability even as the number of objects increases.

3.10. Adaptation to Dynamic Scenes

Handling objects with high speeds and accelerations is a primary motivation for this work. Standard Kalman filters fail in these scenarios because their linear motion model cannot account for abrupt changes. Our ITMC module directly addresses this. By training a transformer on vast amounts of trajectory data, it learns a sophisticated model of object dynamics, including acceleration, deceleration, and sharp turns. When tracking a fast-moving object, the ITMC can predict its future position far more accurately than a linear model, which would significantly lag behind. This inherent ability to model nonlinear dynamics makes KalmanFormer naturally adaptive to challenging, high-speed scenarios without requiring any explicit parameter tuning during runtime.

4. Experiment

4.1. Experiment Settings

We evaluated our method on three well-known benchmark datasets: DanceTrack [10], MOT17 [74], and MOT20 [75]. For evaluation, we use standard tracking metrics: HOTA [76], AssA/DetA [76], IDF1 [74], MOTA [74], and IDS [77]. We train our motion predictor using only the training data from these datasets, without additional samples. For object detection, we use YOLOX [78] detector weights (ByteTrack [29]). The transformer encoder has 6 layers with 8 heads and an embedding dimension of 512. We use Adam [79] with lr = 1×10^{-4} for 50 epochs and a batch size of 64. The history length is $T = 10$, and the masking probability is $p = 0.1$.

Unless otherwise specified, we report HOTA and IDF1 as primary metrics and include AssA/DetA, MOTA, and IDS for completeness.

We primarily report results under the private-detection protocol. Unless otherwise noted, baselines are taken from their respective papers with the authors' recommended detectors. Where detectors differ, we avoid direct claims on detection-driven metrics. We group trackers into appearance-, attention-, correlation/graph-, IoU-, and motion-based families following common practices. Notably, ByteTrack does not exploit appearance features and is treated as IoU-based, and methods such as MOTR/MeMOTR are not categorized as appearance+motion.

4.2. Comparison with State-of-the-Art Methods

We compare KalmanFormer with state-of-the-art trackers on MOT17, DanceTrack and MOT20. As shown in Tables 2–4, KalmanFormer consistently outperforms the strong motion- and IoU-based methods on MOT17 and MOT20 in HOTA and IDF1. On DanceTrack, which exhibits diverse, nonlinear group motions, KalmanFormer remains competitive via robust geometric modeling. Appearance-based trackers often reduce the IDS by leveraging reidentification; our geometry-only design emphasizes motion fidelity and CTAM-based association, narrowing the IDS gap without image features.

To visually compare the two datasets, we provide a line–plot comparison, as shown in Figure 4.

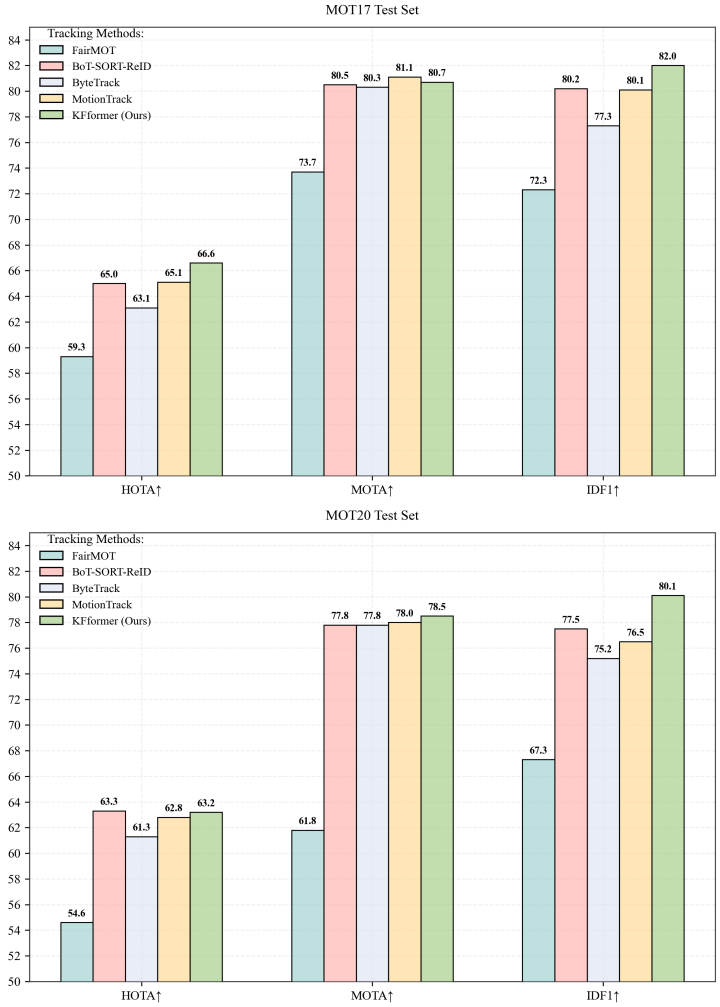


Figure 4. Performance comparison on the MOT17 and MOT20 test sets. The graph displays three key metrics: HOTA (Higher Order Tracking Accuracy, %), IDF1 (ID F1 score, %), and MOTA (Multiple Object Tracking Accuracy, %). Higher values for all three metrics indicate better performance. HOTA measures overall tracking performance, IDF1 evaluates identity preservation, and MOTA focuses on detection accuracy and false positives/negatives. KalmanFormer demonstrates a strong balance across both datasets, particularly excelling in identity preservation (IDF1).

While quantitative metrics demonstrate the overall performance, a qualitative analysis of specific challenging scenarios can better highlight the contributions of our model's core components. Our model shows significant advantages in several key scenarios. In dense crowds with frequent interobject occlusions, such as those found in the MOT20 dataset, the CTAM module models interobject interactions to maintain correct track hypotheses when objects reappear after being occluded, significantly reducing identity switches compared with methods that treat objects in isolation. For highly nonlinear motion, particularly evident in the DanceTrack dataset, where dancers make rapid, unpredictable movements, the ITMC module excels by learning a rich motion model that anticipates turns, accelerations, and decelerations, following complex trajectories more accurately than linear models do. During long-term occlusion and detection failures, our POG module generates reliable pseudo-observations on the basis of learned motion patterns, maintaining track continuity and reducing fragmentation.

We also present qualitative visualizations from MOT17 and MOT20 in Figure 5, illustrating the performance under severe occlusions and dense crowds. The figure is placed near the conclusion for better reading flow.

Table 2. Comparison of KalmanFormer with other methods on the MOT17 test set. The arrows (↑, ↓) indicate whether higher or lower values are better. HOTA and MOTA are measured as percentages (%), IDF1 as a percentage (%), and IDS as a count (the lower percentage is better). Red indicates the best-performing method, and blue indicates the second-best.

Tracker	HOTA↑	MOTA↑	IDF1↑	IDS↓
Appearance-based:				
DAN [80]	39.3	52.4	49.5	8431
QuasiDense [81]	53.9	68.7	66.3	3378
Semi-TCL [82]	59.8	73.3	73.2	2790
FairMOT [11]	59.3	73.7	72.3	3303
CSTrack [83]	59.3	74.9	72.6	3567
ReMOT [84]	59.7	77.0	72.0	2853
StrongSORT [85]	64.4	79.6	79.5	1194
Attention-based:				
CTracker [86]	49.0	66.6	57.4	5529
TransCenter [87]	54.5	73.2	62.2	4614
RelationTrack [88]	61.0	73.8	74.7	1374
TransTrack [25]	54.1	75.2	63.5	3603
MOTRv2 [89]	62.0	78.6	75.0	2619
Graph/Correlation-based:				
GSDT [90]	55.2	73.2	66.5	3891
FUFET [91]	57.9	76.2	68.0	3237
CorrTracker [92]	60.7	76.5	73.6	3369
TransMOT [93]	61.7	76.7	75.1	2346
IoU-based:				
ByteTrack [29]	63.1	80.3	77.3	2196
BoT-SORT [94]	64.6	80.6	79.5	1257
Motion-based:				
TubeTK [95]	48.0	63.0	58.6	4137
CenterTrack [70]	52.2	67.8	64.7	3039
TraDes [96]	52.7	69.1	63.9	3555
MAT [97]	53.8	69.5	63.1	2844
OC-SORT [9]	63.2	78.0	77.5	1950
MotionTrack [98]	65.1	81.1	80.1	1140
KalmanFormer (Ours)	66.6	80.7	82.0	1080

Table 3. Tracking results on the DanceTrack set. The arrows (↑, ↓) indicate whether higher or lower values are better. All metrics (HOTA, IDF1, MOTA, AssA, and DetA) are measured as percentages (%). Red indicates the best-performing method, and blue indicates the second-best.

Tracker	Appear.	Motion	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
Appearance-based:							
DeepSORT [2]	✓	✓	45.6	47.9	87.8	29.7	71.0
ByteTrack [29]	✓	✓	47.3	52.5	89.5	31.4	71.6
MOTR [30]	✓	✓	54.2	51.5	79.7	40.2	73.5
MeMOTR [31]	✓	✓	68.5	71.2	89.9	58.4	80.5
Attention-based:							
TransTrack [25]	✓		45.5	45.2	88.4	27.5	75.9
GTR [99]	✓		48.0	50.3	84.7	31.9	72.5
QuasiDense [81]	✓		54.2	50.4	87.7	36.8	80.1
GHOST [100]	✓		56.7	57.7	91.3	39.8	81.1
Motion-based:							
CenterTrack [70]		✓	41.8	35.7	86.8	22.6	78.1
TraDes [96]		✓	43.3	41.2	86.2	25.4	74.5
SORT [1]		✓	47.9	50.8	91.8	31.2	72.0
OC-SORT [9]		✓	54.6	54.6	89.6	40.2	80.4
KalmanFormer (Ours)		✓	51.44	56.74	88.79	36.90	71.92

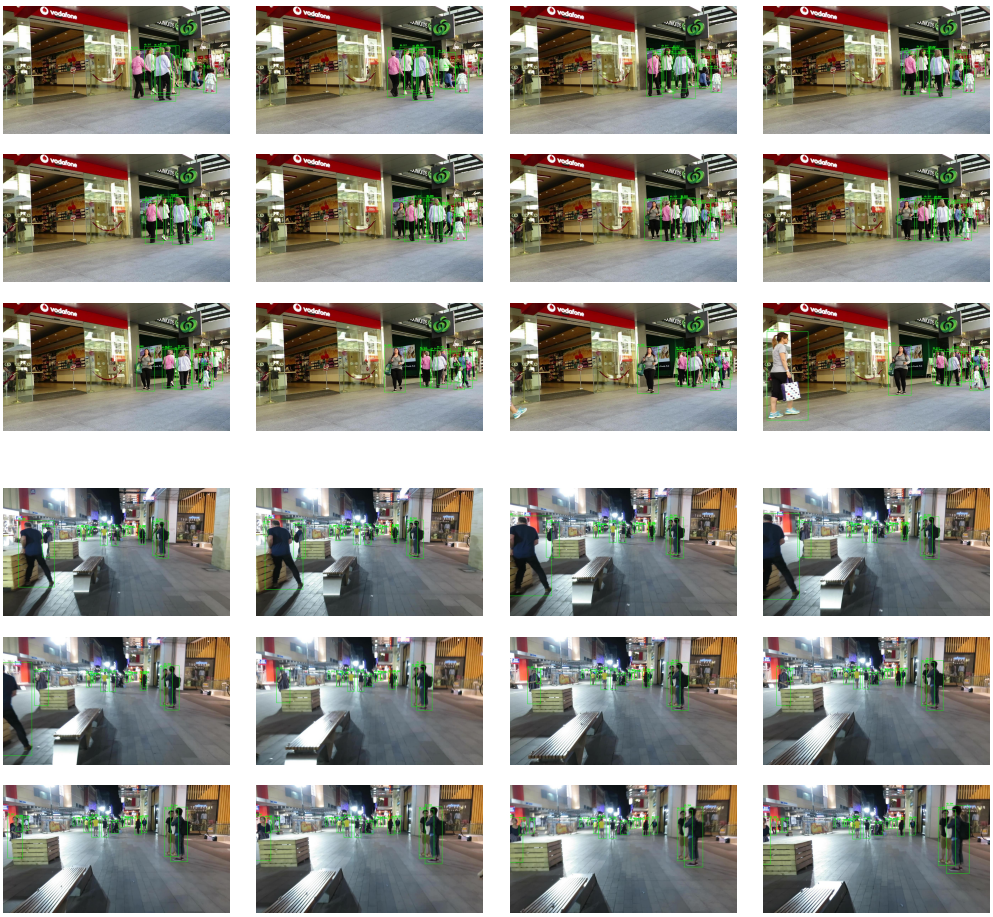


Figure 5. Qualitative visualization of tracking results on MOT17 and MOT20. In the top row (MOT17), KalmanFormer (green box) successfully maintains a consistent ID for the pedestrian emerging from behind the pole, whereas a baseline tracker (green box) incorrectly swaps the ID. This demonstrates our model's superior occlusion handling capability. In the bottom row (MOT20), our method correctly tracks individuals through a dense crowd, maintaining consistent IDs despite multiple overlapping trajectories and partial occlusions. The visualization highlights how our CTAM module helps distinguish between closely interacting pedestrians in crowded scenes with frequent crossings and occlusions.

Table 4. Tracking results on the MOT20 test set under private detection protocols. The arrows (↑, ↓) indicate whether higher or lower values are better. HOTA and MOTA are measured as percentages (%), IDF1 as a percentage (%), and IDS as a count (the lower percentage is better). Red indicates the best-performing method, and blue indicates the second-best.

Tracker	HOTA↑	MOTA↑	IDF1↑	IDS↓
Appearance-based:				
FairMOT [11]	54.6	61.8	67.3	5243
Semi-TCL [82]	55.3	65.2	70.1	4139
CSTrack [83]	54.0	66.6	68.6	3196
UTM [101]	62.5	78.2	76.9	1228
Attention-based:				
TransTrack [25]	48.5	65.0	59.4	3608
RelationTrack [88]	56.5	67.2	70.5	4243
MOTR [30]	57.8	73.4	68.6	2439

Table 4. *Cont.*

Tracker	HOTA↑	MOTA↑	IDF1↑	IDS↓
Graph/Correlation-based:				
MLT [102]	43.2	48.9	54.6	2187
GSDT [90]	53.6	67.1	67.5	3131
Motion-based:				
MotionTrack [98]	62.8	78.0	76.5	1165
ByteTrack [29]	61.3	77.8	75.2	1223
KalmanFormer (Ours)	63.2	78.5	80.1	1930

4.3. Ablation Study

To understand the contribution of each component in KalmanFormer, we conduct an ablation study on the MOT17 validation set. As shown in Table 5, each component contributes complementary gains. ITMC emphasizes long-horizon intratrajectory dynamics, whereas CTAM refines short-horizon interobject interactions; using ITMC alone may slightly reduce association stability when interaction cues are absent, whereas combining CTAM consistently improves all metrics, indicating complementary roles. Regarding hyperparameters, Table 6 suggests that $T = 20$ can be slightly better than $T = 10$ for IDF1 (we choose $T = 10$ as a latency/accuracy trade-off), and Table 7 shows that moderate masking ($p \in [0.1, 0.2]$) improves robustness, whereas large masking harms performance.

Table 5. Ablation study on the MOT17 validation set. **Bold** indicates the best-performing result for each metric.

Components				Metrics		
Kalman Filter	**POG**	**ITMC**	**CTAM**	**HOTA ↑**	**IDF1 ↑**	**MOTA ↑**
✓				67.0	77.8	73.5
✓	✓			67.6	79.2	74.6
✓	✓	✓		65.0	78.1	74.2
✓	✓	✓	✓	**69.14**	**82.7**	**76.2**

Table 6. Impact of historical trajectory embedding length T. The arrows (↑, ↓) indicate whether higher or lower values are better. **Bold** indicates the parameter value that achieves the best performance.

	IDF1↑					
T	5	10	20	30	40	50
KalmanFormer (Ours)	75.2	78.1	**82.7**	80.4	77.6	77.8

Table 7. Impact of masked tokens at varying probabilities p. The arrows (↑, ↓) indicate whether higher or lower values are better. **Bold** indicates the parameter value that achieves the best performance.

	IDF1↑					
p	0	0.05	0.1	0.2	0.3	0.4
KalmanFormer (Ours)	76.1	78.2	80.4	**82.7**	73.6	40.2

5. Conclusions

In this study, a hybrid multi-object tracking framework called KalmanFormer was developed by integrating a deep motion model into the SORT tracking pipeline. Critical limitations in existing tracking methods include three key innovations: (1) an inner-trajectory motion corrector (ITMC) that captures nonlinear motion patterns from historical trajectories, (2) a cross-trajectory attention module (CTAM) that models interobject interactions to improve associations under occlusions, and (3) a pseudo-observation generator (POG) that provides reliable neural-based predictions when detections are missing.

Comprehensive experiments were conducted on the MOT17, MOT20, and DanceTrack benchmarks. The results demonstrate that KalmanFormer achieves state-of-the-art performance, with HOTA scores of 66.6% on MOT17 and 63.2% on MOT20. The model is particularly effective at reducing identity switches, achieving an IDF1 score of 82.0% on MOT17, which represents a significant improvement over previous approaches. These results validate the hypothesis that integrating transformer-based components with traditional Kalman filtering can substantially enhance tracking performance in challenging scenarios.

Several important insights were revealed through the experimental results. First, the combination of ITMC and CTAM provides complementary benefits, with ITMC handling nonlinear motion and CTAM addressing interobject interactions. Second, the sparse attention mechanism ensures that the computational complexity scales linearly with the number of objects, making the approach suitable for real-time applications. Third, the pseudo-observation mechanism significantly improves tracking continuity during occlusions, preventing the drift that typically affects Kalman-filter-based trackers.

Despite these advantages, the approach has limitations that warrant further investigation. The primary limitation is the reliance on geometric information alone, which can be insufficient for reidentification after very long occlusions or in cases of extreme visual ambiguity between two objects. An appearance-based method such as StrongSORT may be more effective at reidentifying a person who disappears for hundreds of frames and reappears elsewhere in the scene. Additionally, while the model generalizes well to the tested datasets, its performance on highly specialized domains with unique motion patterns might require domain-specific fine-tuning.

Several promising directions for future work emerge from this research:

- Appearance Integration: A hybrid model that combines this motion-based approach with lightweight appearance features could further improve reidentification after prolonged occlusions. KalmanFormer's advanced motion predictions can be fused with a lightweight, efficient appearance feature extractor (e.g., MobileNet). The motion model would handle short-term tracking and occlusions, whereas the appearance features would only be invoked for resolving ambiguities or for reidentification after prolonged track loss. This preserves most of the model's efficiency while adding another layer of robustness.
- Online Adaptation: Mechanisms for online adaptation of the transformer components could be incorporated to enhance performance in domains with evolving motion patterns.
- Multimodal Fusion: The framework could be extended to incorporate multiple sensor modalities (e.g., LiDAR and radar) to improve robustness under adverse environmental conditions.
- Extended Dataset Evaluation: While the current work focuses on the MOT17, MOT20, and DanceTrack datasets, future work should evaluate the approach on a wider range of datasets, such as SportsMOT and BDD100K, to better understand the model's performance across diverse scenarios, including sports analytics and autonomous driving applications.

In conclusion, the integration of classical filtering techniques with deep learning approaches for multi-object tracking has been demonstrated through the KalmanFormer framework. The system handles motion patterns and occlusions while maintaining computational efficiency, offering potential benefits for various tracking applications.

Author Contributions: Conceptualization, X.W. and Y.Q.; methodology, J.H., Y.L., X.W., and Y.Q.; software, J.Y. and Y.Q.; validation, Y.Q.; formal analysis, X.W.; investigation, Y.L., J.Y. and Y.Q.; resources, J.H., Y.L., and W.X.; data curation, J.H. and W.X.; writing—original draft preparation, J.H.

and Y.L.; writing—review and editing, J.Y., X.W., W.X., and Y.Q.; visualization, J.Y., X.W., and W.X.; supervision, X.W.; project administration, X.W.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by the Chongqing New YC Project under Grant CSTB2024YCJH-KYXM0126; the Postdoctoral Fellowship Program of CPSF under Grant GZC20233322, and the Postdoctoral Talent Special Program; the General Program of the Natural Science Foundation of Chongqing under Grant CSTB2024NSCQ-MSX0479; Chongqing Postdoctoral Foundation Special Support Program under Grant 2023CQBSHTB3119; China Postdoctoral Science Foundation under Grant 2024MD754244. W.X. was supported by Grant CSTB2024YCJH-KYXM0126; Y.J. was supported by Grant GZC20233322 and the Postdoctoral Talent Special Program; and X.W. was supported by Grants CSTB2024NSCQ-MSX0479, 2023CQBSHTB3119 and 2024MD754244.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upadhya, A. Simple Online and Realtime Tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468. [CrossRef]
2. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 3645–3649. [CrossRef]
3. Kalman, R.E. *A New Approach To Linear Filtering and Prediction Problems*; Wiley Press: New York, NY, USA, 2001; Volume 82D, pp. 35–45. [CrossRef]
4. Zhou, M.; Li, J.; Wei, X.; Luo, J.; Pu, H.; Wang, W.; He, J.; Shang, Z. AFES: Attention-Based Feature Excitation and Sorting for Action Recognition. *IEEE Trans. Consum. Electron.* **2025**, *71*, 5752–5760. [CrossRef]
5. Shen, W.; Zhou, M.; Chen, Y.; Wei, X.; Feng, Y.; Pu, H.; Jia, W. Image Quality Assessment: Investigating Causal Perceptual Effects with Abductive Counterfactual Inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 17990–17999. [CrossRef]
6. Lan, X.; Xian, W.; Zhou, M.; Yan, J.; Wei, X.; Luo, J.; Jia, W.; Kwong, S. No-Reference Image Quality Assessment: Exploring Intrinsic Distortion Characteristics via Generative Noise Estimation with Mamba. *IEEE Trans. Circuits Syst. Video Technol.* **2025**. [CrossRef]
7. Zheng, Z.; Zhou, M.; Shang, Z.; Wei, X.; Pu, H.; Luo, J.; Jia, W. GAANet: Graph Aggregation Alignment Feature Fusion for Multispectral Object Detection. *IEEE Trans. Ind. Inform.* **2025**. [CrossRef]
8. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. *arXiv* **2021**, arXiv:2104.13840. [CrossRef]
9. Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; Kitani, K. Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 9686–9696. [CrossRef]
10. Sun, Y.; Zhang, W.; Zhao, B.; Li, L.; Wang, J. DanceTrack: Robust Multi-Object Tracking with Motion-Aware Object Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022. [CrossRef]
11. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
12. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
13. Julier, S.J.; Uhlmann, J.K. New extension of the Kalman filter to nonlinear systems. In Proceedings of the Signal Processing, Sensor Fusion, and Target Recognition VI, SPIE, Orlando, FL, USA, 21–25 April 1997; Volume 3068, pp. 182–193. [CrossRef]
14. Gustafsson, F.; Gunnarsson, F.; Bergman, N.; Forssell, U.; Jansson, J.; Karlsson, R.; Nordlund, P.J. Particle filters for positioning, navigation, and tracking. *IEEE Trans. Signal Process.* **2002**, *50*, 425–437. [CrossRef]
15. Han, Q.; Fan, Z.; Dai, Q.; Sun, L.; Cheng, M.M.; Liu, J.; Wang, J. On the connection between local attention and dynamic depth-wise convolution. *arXiv* **2022**, arXiv:2106.04263. [CrossRef]

16. Zhou, M.; Wei, X.; Wang, S.; Kwong, S.; Fong, C.; Wong, P.; Yuen, W. Global rate-distortion optimization-based rate control for HEVC HDR coding. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4648–4662. [CrossRef]
17. Zhou, M.; Zhang, Y.; Li, B.; Lin, X. Complexity correlation-based CTU-level rate control with direction selection for HEVC. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**, *13*, 1–23. [CrossRef]
18. Zhou, M.; Wei, X.; Ji, C.; Xiang, T.; Fang, B. Optimum quality control algorithm for versatile video coding. *IEEE Trans. Broadcast.* **2022**, *68*, 582–593. [CrossRef]
19. Wei, X.; Zhou, M.; Wang, H.; Yang, H.; Chen, L.; Kwong, S. Recent advances in rate control: From optimization to implementation and beyond. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *34*, 17–33. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Red Hook, NY, USA, 4–9 December 2017; pp. 6000–6010. [CrossRef]
21. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. TrackFormer: Multi-Object Tracking with Transformers. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022. [CrossRef]
22. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 12934–12949. [CrossRef]
23. Shen, Y.; Feng, Y.; Fang, B.; Zhou, M.; Kwong, S.; Qiang, B. DSRPH: Deep semantic-aware ranking preserving hashing for efficient multi-label image retrieval. *Inf. Sci.* **2022**, *539*, 145–156. [CrossRef]
24. Gao, T.; Sheng, W.; Zhou, M.; Fang, B.; Luo, F.; Li, J. Method for fault diagnosis of temperature-related mems inertial sensors by combining Hilbert-Huang transform and deep learning. *Sensors* **2020**, *20*, 5633. [CrossRef]
25. Sun, P.; Jiang, Y.; Zhang, R.; Xie, E.; Cao, J.; Hu, X.; Kong, T.; Yuan, Z.; Wang, C.; Luo, P. Transtrack: Multiple-object tracking with transformer. *arXiv* **2020**, arXiv:2012.15460. [CrossRef]
26. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578. [CrossRef]
27. Zhou, M.; Li, Y.; Yang, G.; Wei, X.; Pu, H.; Luo, J. COFNet: Contrastive Object-aware Fusion using Box-level Masks for Multispectral Object Detection. *IEEE Trans. Multimed.* **2025**. [CrossRef]
28. Zhou, M.; Zhao, X.; Luo, F.; Luo, J.; Pu, H.; Xiang, T. Robust RGB-T Tracking via Adaptive Modality Weight Correlation Filters and Cross-modality Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* **2023**, *20*, 1–20. [CrossRef]
29. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–21. [CrossRef]
30. Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; Wei, Y. MOTR: End-to-End Multiple-Object Tracking with TRansformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2022. [CrossRef]
31. Gao, J.; Zhang, H.; Liu, Q.; Li, W.; Zhang, L.; Li, X. MeMOTR: Memory-augmented Multi-Object Tracking with Transformer. *arXiv* **2023**, arXiv:2301.03482. [CrossRef]
32. Henriques, J.F.; Caseiro, R.; Pereira, F.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 583–596. [CrossRef]
33. Zhang, S.; Shang, Z.; Zhou, M.; Wang, Y.; Sun, G. Cross-modal identity correlation mining for visible-thermal person re-identification. *Multimed. Tools Appl.* **2022**, *81*, 39981–39994. [CrossRef]
34. Song, J.; Zhou, M.; Luo, J.; Pu, H.; Feng, Y.; Wei, X.; Jia, W. Boundary-Aware Feature Fusion With Dual-Stream Attention for Remote Sensing Small Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5600213. [CrossRef]
35. Luo, F.; Zhou, M.; Fang, B. Correlation filters based on strong spatio-temporal for robust RGB-T tracking. *J. Circuits Syst. Comput.* **2022**, *31*, 2250041. [CrossRef]
36. Cheng, S.; Song, J.; Zhou, M.; Wei, X.; Pu, H.; Luo, J.; Jia, W. EF-DETR: A Lightweight Transformer-Based Object Detector With an Encoder-Free Neck. *IEEE Trans. Ind. Inform.* **2024**, *20*, 12994–13002. [CrossRef]
37. Shen, W.; Zhou, M.; Luo, J.; Li, Z.; Kwong, S. Graph-represented distribution similarity index for full-reference image quality assessment. *IEEE Trans. Image Process.* **2023**, *33*, 3075–3089. [CrossRef]
38. Huang, H.; Liang, Y.; Tsoi, A.C.; Lo, S.L.; Leung, A.P. A novel bagged particle filter for object tracking. In Proceedings of the 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum, Zhuhai, China, 3–4 December 2016; pp. 331–338. [CrossRef]
39. Yan, J.; Zhang, B.; Zhou, M.; Kwok, H.; Siu, S. Multi-Branch-CNN: Classification of ion channel interacting peptides using multi-branch convolutional neural network. *Comput. Biol. Med.* **2022**, *147*, 105717. [CrossRef] [PubMed]
40. Wei, X.; Li, J.; Zhou, M.; Wan, X. Contrastive distortion-level learning-based no-reference image-quality assessment. *Int. J. Intell. Syst.* **2022**, *37*, 8730–8746. [CrossRef]

41. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards real-time multi-object tracking. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 107–122. [CrossRef]
42. Chen, L.; Ai, H.; Zhuang, Z.; Shang, C. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6. [CrossRef]
43. Held, D.; Thrun, S.; Savarese, S. Learning to Track at 100 FPS with Deep Regression Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2016; pp. 749–765. [CrossRef]
44. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*; Springer: Cham, Switzerland, 2016; pp. 850–865. [CrossRef]
45. Bochinski, E.; Eiselein, V.; Sikora, T. High-speed tracking-by-detection without using image information. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6. [CrossRef]
46. Cui, Y.; Guo, C.; Yin, F.; Liu, X.; Li, B.; Liu, H.; Wang, L. SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 367–383. [CrossRef]
47. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Paris, France, 1–6 October 2023; pp. 2636–2645. [CrossRef]
48. Li, W.; Meng, W.; Li, B.; Zhang, J.; Zhang, X. SCOOT: Self-supervised CentBewleyric Open-set Object Tracking. In Proceedings of the SIGGRAPH Asia 2023 Posters, Sydney, Australia, 12–15 December 2023. [CrossRef]
49. Liao, X.; Wei, X.; Zhou, M.; Li, Z.; Kwong, S. Image Quality Assessment: Measuring Perceptual Degradation via Distribution Measures in Deep Feature Spaces. *IEEE Trans. Image Process.* **2024**, *33*, 4044–4059. [CrossRef] [PubMed]
50. Li, W.; Meng, W.; Li, B.; Zhang, J.; Zhang, X. DMA-YOLO: Multi-scale object detection method with attention mechanism for aerial images. *Vis. Comput.* **2023**, *40*, 4505–4518. [CrossRef]
51. Zhou, M.; Wang, H.; Wei, X.; Feng, Y.; Luo, J.; Pu, H.; Zhao, J.; Wang, L.; Chu, Z.; Wang, X.; et al. HDIQA: A Hyper Debiasing Framework for Full Reference Image Quality Assessment. *IEEE Trans. Broadcast.* **2024**, *70*, 545–554. [CrossRef]
52. Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; Felsberg, M. ATOM: Accurate Tracking by Overcoming Drift. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4660–4669. [CrossRef]
53. Guo, W.; Zhang, Y.; Li, W.; Wang, J.; Zhang, X. Community-based social recommendation under local differential privacy protection. *Inf. Sci.* **2023**, *639*, 119002. [CrossRef]
54. Zhao, W.; Zhang, Y.; Li, W.; Wang, J.; Zhang, X. Siamese Networks: The Tale of Two Manifolds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
55. Wei, X.; Zhou, M.; Jia, W. Toward Low-Latency and High-Quality Adaptive 360° Streaming. *IEEE Trans. Ind. Inform.* **2023**, *19*, 6326–6336. [CrossRef]
56. Li, B.; Wu, W.; Zhang, X.; Yan, J.; Xu, D.; Li, H. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6518–6527. [CrossRef]
57. Lang, S.; Liu, X.; Zhou, M.; Luo, J.; Pu, H.; Zhuang, X.; Wang, J.; Wei, X.; Zhang, T.; Feng, Y.; et al. A Full-Reference Image Quality Assessment Method via Deep Meta-Learning and Conformer. *IEEE Trans. Broadcast.* **2024**, *70*, 316–324. [CrossRef]
58. Liao, X.; Wei, X.; Zhou, M.; Kwong, S. Full-Reference Image Quality Assessment: Addressing Content Misalignment Issue by Comparing Order Statistics of Deep Features. *IEEE Trans. Broadcast.* **2024**, *70*, 305–315. [CrossRef]
59. Zhou, M.; Zhang, Y.; Li, B.; Hu, H.M. Complexity-based intra frame rate control by jointing inter-frame correlation for high efficiency video coding. *J. Vis. Commun. Image Represent.* **2017**, *42*, 46–64. [CrossRef]
60. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8126–8135. [CrossRef]
61. Zhou, M.; Hu, H.M.; Zhang, Y. Region-based intra-frame rate-control scheme for high efficiency video coding. In Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, Siem Reap, Cambodia, 9–12 December 2014; pp. 1–4. [CrossRef]
62. Zhou, M.; Leng, H.; Fang, B.; Xiang, T.; Wei, X.; Jia, W. Low-light image enhancement via a frequency-based model with structure and texture decomposition. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–23. [CrossRef]
63. Shen, W.; Zhou, M.; Wei, X.; Wang, H.; Fang, B.; Ji, C.; Zhuang, X.; Wang, J.; Luo, J.; Pu, H.; et al. A Blind Video Quality Assessment Method via Spatiotemporal Pyramid Attention. *IEEE Trans. Broadcast.* **2024**, *70*, 251–264. [CrossRef]

64. Xian, W.; Zhou, M.; Fang, B.; Xiang, T.; Jia, W.; Chen, B. Perceptual Quality Analysis in Deep Domains Using Structure Separation and High-Order Moments. *IEEE Trans. Multimed.* **2024**, *26*, 2219–2234. [CrossRef]
65. Duan, C.; Feng, Y.; Zhou, M.; Xiong, X.; Wang, Y.; Qiang, B.; Jia, W. Multilevel Similarity-Aware Deep Metric Learning for Fine-Grained Image Retrieval. *IEEE Trans. Ind. Inform.* **2023**, *19*, 9173–9182. [CrossRef]
66. Yan, J.; Zhou, M.; Pan, J.; Yin, M.; Fang, B. Recent advances in 3d human pose estimation: From optimization to implementation and beyond. *INternational J. Pattern Recognit. Artif. Intell.* **2022**, *36*, 2255003. [CrossRef]
67. Liu, T.; Lin, X.; Jia, W.; Zhou, M.; Zhao, W. Regularized attentive capsule network for overlapped relation extraction. *arXiv* **2020**, arXiv:2012.10187. [CrossRef]
68. Wang, G.; Zhang, Y.; Li, B.; Fan, R.; Zhou, M. A fast and HEVC-compatible perceptual video coding scheme using a transform-domain Multi-Channel JND model. *Multimed. Tools Appl.* **2018**, *77*, 12777–12803. [CrossRef]
69. Zhou, M.; Li, B.; Zhang, Y. Content-adaptive parameters estimation for multi-dimensional rate control. *J. Vis. Commun. Image Represent.* **2016**, *34*, 204–218. [CrossRef]
70. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking Objects as Points. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020. [CrossRef]
71. Zhao, Y.; Liao, X.; He, X.; Zhou, M.; Li, C. Accelerated Primal-Dual Mirror Dynamics for Centralized and Distributed Constrained Convex Optimization Problems. *J. Mach. Learn. Res.* **2023**, *24*, 1–59. [CrossRef]
72. Hu, H.M.; Zhou, M.; Liu, Y.; Yin, N. A region-based intra-frame rate control scheme by jointing inter-frame dependency and inter-frame correlation. *Multimed. Tools Appl.* **2017**, *76*, 12917–12940. [CrossRef]
73. Gan, Y.; Xiang, T.; Liu, H.; Ye, M.; Zhou, M. Generative adversarial networks with adaptive learning strategy for noise-to-image synthesis. *Neural Comput. Appl.* **2023**, *35*, 6197–6206. [CrossRef]
74. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831. [CrossRef]
75. Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; Leal-Taixé, L. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv* **2020**, arXiv:2003.09003. [CrossRef]
76. Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; Leibe, B. Hota: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 548–578. [CrossRef]
77. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 17–35. [CrossRef]
78. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
79. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]
80. Sun, S.; Akhtar, N.; Song, H.; Mian, A.; Shah, M. DAN: Deep Affinity Network for Multiple Object Tracking. *IEEE Trans. Image Process.* **2020**, *29*, 2741–2755. [CrossRef]
81. Pang, J.; Qiu, L.; Li, X.; Chen, H.; Li, Q.; Darrell, T.; Yu, F. Quasi-Dense Similarity Learning for Multiple Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021. [CrossRef]
82. Li, W.; Xiong, Y.; Yang, S.; Xu, M.; Wang, Y.; Xia, W. Semi-TCL: Semi-Supervised Track Contrastive Representation Learning. *arXiv* **2021**, arXiv:2107.02396. [CrossRef]
83. Liang, C.; Zhang, Z.; Zhou, X.; Li, B.; Zhu, S.; Hu, W. Rethinking the Competition Between Detection and ReID in Multiobject Tracking. *IEEE Trans. Image Process.* **2022**, 3182–3196. [CrossRef]
84. Yang, F.; Chang, X.; Sakti, S.; Wu, Y.; Nakamura, S. ReMOT: A model-agnostic refinement for multiple object tracking. *Image Vis. Comput.* **2021**, *106*, 104091. [CrossRef]
85. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. StrongSORT: Make DeepSORT Great Again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [CrossRef]
86. Peng, J.; Wang, C.; Wan, F.; Wu, Y.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Fu, Y. Chained-Tracker: Chaining Paired Attentive Regression Results for End-to-End Joint Multiple-Object Detection and Tracking. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020. [CrossRef]
87. Xu, Y.; Ban, Y.; Delorme, G.; Gan, C.; Rus, D.; Alameda-Pineda, X. TransCenter: Transformers With Dense Representations for Multiple-Object Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 7820–7835. [CrossRef] [PubMed]
88. Yu, E.; Li, Z.; Han, S.; Wang, H. RelationTrack: Relation-aware Multiple Object Tracking with Decoupled Representation. *arXiv* **2021**, arXiv:2105.04322. [CrossRef]
89. Zhang, Y.; Wang, T.; Zhang, X. MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors. In Proceedings of the Computer Vision and Pattern Recognition CVPR, Vancouver, BC, Canada, 17–24 June 2023. [CrossRef]
90. Wang, Y.; Kitani, K.; Weng, X. Joint object detection and multi-object tracking with graph neural networks. *arXiv* **2020**, arXiv:2006.13164. [CrossRef]

91. Shan, C.; Wei, C.; Deng, B.; Huang, J.; Hua, X.S.; Cheng, X.; Liang, K. Tracklets predicting based adaptive graph tracking. *arXiv* **2020**, arXiv:2010.09015. [CrossRef]
92. Wang, Q.; Zheng, Y.; Pan, P.; Xu, Y. Multiple Object Tracking with Correlation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3876–3886. [CrossRef]
93. Chu, P.; Wang, J.; You, Q.; Ling, H.; Liu, Z. TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking. *arXiv* **2021**, arXiv:2104.00194. [CrossRef]
94. Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv* **2022**, arXiv:2206.14651. [CrossRef]
95. Pang, B.; Li, Y.; Zhang, Y.; Li, M.; Lu, C. TubeTK: Adopting Tubes to Track Multi-Object in a One-Step Training Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6308–6318. [CrossRef]
96. Wu, J.; Cao, J.; Song, L.; Wang, Y.; Yang, M.; Yuan, J. Track to Detect and Segment: An Online Multi-Object Tracker. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12352–12361. [CrossRef]
97. Han, S.; Huang, P.; Wang, H.; Yu, E.; Liu, D.; Pan, X.; Zhao, J. MAT: Motion-Aware Multi-Object Tracking. *Neural Comput. Appl.* **2021**, *476*, 75–86. [CrossRef]
98. Qin, Z.; Zhou, S.; Wang, L.; Duan, J.; Hua, G.; Tang, W. MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking. *arXiv* **2023**, arXiv:2303.10404. [CrossRef]
99. Zhou, X.; Yin, T.; Koltun, V.; Krähenbühl, P. Global tracking transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8771–8780. [CrossRef]
100. Seidenschwarz, J.; Brasó, G.; Serrano, V.C.; Elezi, I.; Leal-Taixé, L. Simple Cues Lead to a Strong Multi-Object Tracker. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13813–13823. [CrossRef]
101. You, S.; Yao, H.; Bao, B.K.; Xu, C. UTM: A Unified Multiple Object Tracking Model With Identity-Aware Feature Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023. [CrossRef]
102. Zhang, Y.; Sheng, H.; Wu, Y.; Wang, S.; Ke, W.; Xiong, Z. Multiplex labeling graph for near-online tracking in crowded scenes. *IEEE Internet Things J.* **2020**, *7*, 7892–7902. [CrossRef]

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