Temporal Coherent Object Flow for Multi-Object Tracking

Zikai Song[1], Run Luo[2], Lintao Ma[1], Ying Tang[1], Yi-Ping Phoebe Chen[3], Junqing Yu[1], Wei Yang[1]*

[1]Huazhong University of Science and Technology, Wuhan, China
[2]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[3]La Trobe University, Melbourne, Australia
{skyesong, mltdml, t_ying, yjqing, weiyangcs}@hust.edu.cn, r.luo@siat.ac.cn, phoebe.chen@latrobe.edu.cn

Abstract

Multi-object tracking is a challenging vision task that requires simultaneous reasoning about object detection and object association. Conventional solutions use frame as the basic unit and typically rely on a motion predictor that exploits the appearance features to associate detected candidates, leading to insufficient adaptability to long-term associations. In this study, we propose a section-based multi-object tracking approach that integrates a temporal coherent **O**bject **F**low **Track**er (OFTrack), capable of achieving simultaneous multi-frame tracking by treating multiple consecutive frames as the basic processing unit, denoted as a "section". Our OFTrack boosts the optical flow to the object flow by employing object perception and section-based motion estimation strategies. Object perception adopts object-aware sampling and scale-aware correlation to enable precise target discrimination. Motion estimation models the correlation of different objects in multi-frames via specialized temporal-spatial attention to achieve robust association in very long videos. Additionally, to address the oscillation of unpredictable trajectories in multi-frame estimation, we have designed temporal coherent enhancement including the trajectory masking pretraining and the smoothing constraint on trajectory curves. Comprehensive experiments on several widely used benchmarks demonstrate the superior performance of our approach.

Introduction

Multi-object tracking (MOT) is a challenging vision problem and has many real-world applications (Zhou, Yu, and Yang 2023; Song et al. 2022, 2024; Ye et al. 2024), such as video surveillance, autonomous driving, and etc. The primary objective of MOT is to identify and track numerous objects of interest in dynamic scenes and to maintain their identities across successive frames. The challenge stems from inherent complexities (Song et al. 2023), including the intricate association in crowded scenes, the visual resemblance between targets, and complex motion patterns.

To address the problems, existing MOT approaches predominantly adhere to two distinct strategies: the two-stage tracking-by-detection methods and the one-stage object query network methods. **Two-stage** tracking-by-detection paradigm (Bewley et al. 2016) first detects objects in individual frames using detectors and then relies on a motion tracker to associate the detected objects between consecutive frames. Typically, the motion tracker is a Kalman filter (Bishop, Welch et al. 2001) or its modifications, which predicts the trajectories based on previous states. It performs well for smooth and linear motion patterns, but can easily fail in the presence of complex motions and object scale change. **One-stage** object query network paradigm (Zeng et al. 2022) is a recent research trend that mainly extend the DETR (Carion et al. 2020) for MOT. They adopt the query-based scheme, represent each target object as a query, and force to regress the same instance across frames. These approaches perform implicit inter-frame target association, discard the motion process and can not inherit well from the skills of motion operations (Zhang et al. 2022), culminating in diminished association ability and inferior performances.

In this work, we introduce a new-design temporal coherent object flow tracker (OFTrack), that uses multiple con-

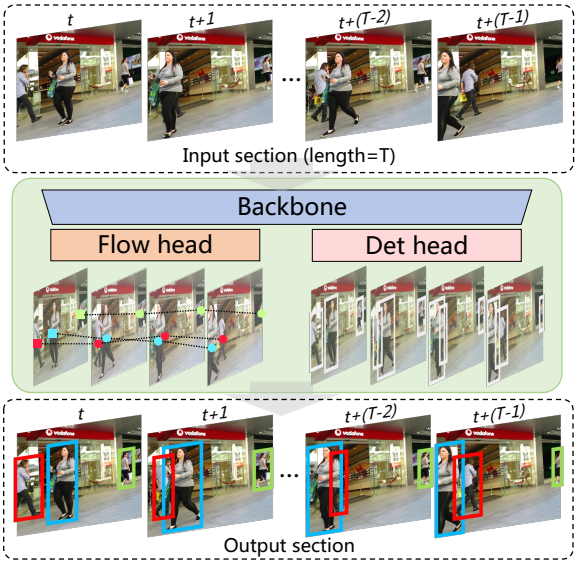


Figure 1: OFTrack jointly reasons detection and tracking through a detection head (det head) and a motion estimation head (flow head). The flow head shares the same backbone with the det head and enables simultaneous tracking of target objects across multiple frames within a section.

*indicates corresponding author.

secutive frames as the basic processing unit, i.e., tracking objects in multiple frames within the section simultaneously. OFTrack exploits an object flow network to provide temporal coherent motion estimation, ensuring robust tracking in long sequences with large intervals. Additionally, to address the challenges posed by network learning for simultaneous multi-frame estimation, we have devised a temporal coherent enhancement strategy. Which involves a pre-training with masked trajectories and a smoothing constraints on trajectory curves, leading to better motion estimation.

Our flow head is inspired by the pixel-level optical flow. However, boosting the optical flow to object flow presents several challenges, object flow in tracking tasks requires semantic awareness of the objects rather than pixels, and needs to estimate the motion of objects over an extended period rather than only at an infinitesimal distance in the optical flow. To tackle these challenges, we propose object perception and motion estimation that enhance semantic awareness and adaptability for long-term association. In object perception, we design the object-aware sampling strategy to accurately sample the features centered on tracked targets and their surrounding features. We also employ the scale-aware correlation which uses bidirectional multi-scale processing to generate correlation volumes. In motion estimation, we propose spatial-temporal attention to simultaneously predict the object's motion in multiple frames within a section. Spatial attention interacts the correlation of different objects within a frame, and temporal attention considers the same object across frames for the duration. Additionally, to avoid the oscillation and unreliability during the estimation of long trajectories, we adopt the bidirectional pre-train to estimate the trajectories through randomly masking tracks, and the curvature constraint which utilizes smoothing constraints of trajectory curve to avoid oscillation and uncertainty in estimation of long trajectory. Extensive experiments on several challenging datasets such as MOT17 (Milan et al. 2016), MOT20 (Dendorfer et al. 2020), DanceTrack (Sun et al. 2022) and KITTI (Geiger, Lenz, and Urtasun 2012) exhibit state-of-the-art performance of our approach.

In summary, our main contributions include:

1. An multi-object tracking framework, which attaches an object flow network to the detector, achieves the jointly reasoning for both object detection and tracking.
2. An object flow network, in which the object perception and motion estimation strategies are introduced to enhance semantic awareness and adaptability for long-term association by simultaneously predicting the object motion in multiple frames.
3. A temporal coherent enhancement strategy, in which the bidirectional pre-train and the curvature constraint are designed to achieve stronger estimation capabilities during the estimation of long-term trajectories.

Related Work

Two-Stage Tracking by Detection Methods

One of the predominant schemes of multiple object trackers is the *tracking-by-detection paradigm* (Bewley et al. 2016; Kumar, Charpiat, and Thonnat 2015). They first predict the object bounding boxes through object detectors (Ren et al. 2015; Ge et al. 2021) for each frame, and then associate the detected objects using a separate motion tracker between consecutive frames. SORT (Bewley et al. 2016) first introduces the Kalman filter to track objects and associates each bounding box with its highest overlapping by the Hungarian algorithm (Kuhn 1955). DeepSORT (Wojke, Bewley, and Paulus 2017) improves the association in the SORT with motion and deep appearance features. ByteTrack (Zhang et al. 2022) utilizes the similarities of low confidence detections to tackle the problem of non-negligible missing detection and fragmented trajectories. P3AFormer (Zhao et al. 2022) adopts pixel-wise distribution architecture and combines with the Kalman filter to enhance the object association. OC-SORT (Cao et al. 2022) improves the linear motion assumption in the Kalman filter for better adapting the occlusion and non-linear motion. MotionTrack (Qin et al. 2023) address the short-range and long-range association problems by modeling all interactions between targets and re-identifying the lost targets through correlation calculation and error compensation. SUSHI (Cetintas, Brasó, and Leal-Taixé 2023) processes long clips by splitting them into a hierarchy of sub-clips, uses a graph neural network for the association of two adjacent clips as in TrackMPNN(Rangesh et al. 2021), to generate increasingly longer trajectories at every level of hierarchy. SparseTrack (Liu et al. 2023) leverages the pseudo-depth method to estimate the relative depth relationship between different targets and divides the target set into multiple sparse subsets in order of increasing depth.

Unlike the two-stage approach of separately conducting object detection and association, our method achieves multi-object tracking more efficiently by utilizing a joint model for object flow motion and detection.

One-Stage Methods

The one-stage paradigm has been made great explorations in recent years, which joint detection and association pipeline and aims to convert detectors into trackers to achieve detection and tracking simultaneously in a single stage.

Query-based methods are a recent research trend (Zhang, Wang, and Zhang 2023; Lv et al. 2024) that mainly extends the DETR (Zhu et al. 2020) for MOT. These methods represent each target object as a query and force it to regress the same instance across different frames. TrackFormer (Meinhardt et al. 2022) and MOTR (Zeng et al. 2022) concatenate the object and auto-regressive track query to perform object detection and association simultaneously. TransTrack (Sun et al. 2020) passes track features cyclically to learn the aggregated embedding of each object. MeMOT (Cai et al. 2022) preserves a large spatio-temporal memory and uses an attention aggregator to encode past observations. Query-based approaches perform implicit inter-frame target association and offer an end-to-end integrated computational process for the entire model. GTR (Zhou et al. 2022) associates objects across all frames of the input video clip, it encodes detections from multiple consecutive frames, and uses trajectory queries to group them into trajectories. DiffusionTrack (Luo et al. 2023) formulate MOT as a generative

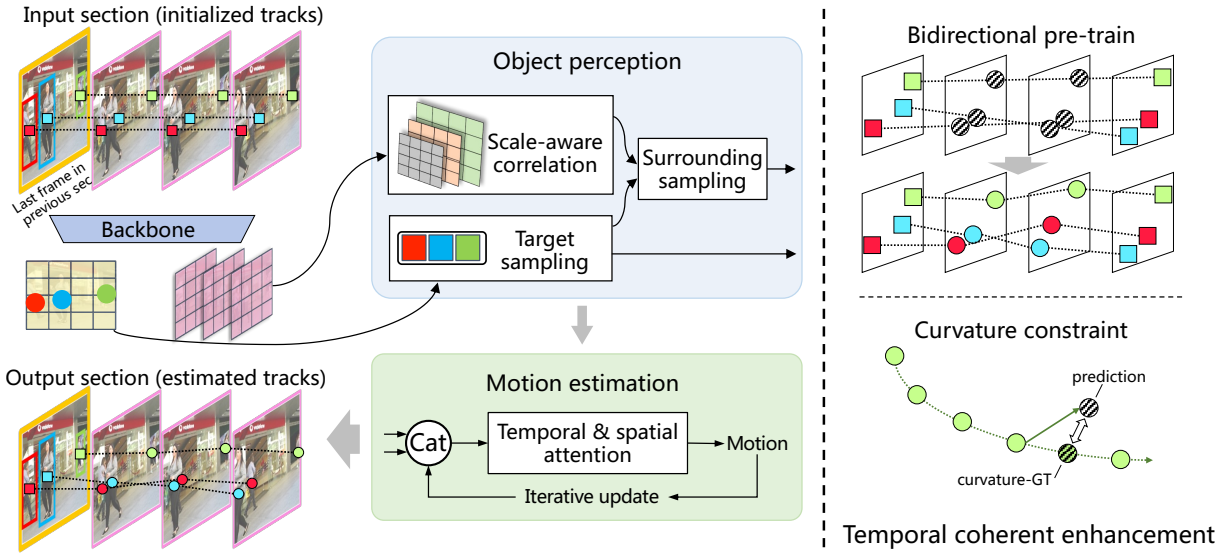


Figure 2: The pipeline of flow head (left) and the temporal coherent enhancement (right). The **flow head** comprises the object perception and the motion estimation. We extract features from all frames within a section and sample object queries from the first frame. Then, we perform scale-aware correlation and sample surrounding features. Finally, our specialized spatial-temporal attention iteratively updates the object motion. The **temporal coherent enhancement** comprises a bidirectional pre-train and a curvature constraint, enhance the estimation capability of the flow head during training by employing trajectory masking pre-training and a smoothing constraint loss function.

noise-to-tracking diffusion process, which refines the coordinates of all paired boxes in a coarse-to-fine paradigm to associate same object.

However, due to the absence of explicit position changes, query-based approaches are unable to extend excellent works based on motion operations (Teed and Deng 2020). Additionally, a fixed number of queries makes it challenging to detect objects in complex and crowded visual scenes.

Tracking-by-regression methods avoid the association between frames, opting instead to achieve tracking by regressing past object locations to their new positions. CenterTrack (Zhou, Koltun, and Krähenbühl 2020) uses tracking-conditioned detection to localize objects and predict their offsets. In MPNTrack (Brasó and Leal-Taixé 2020), a graph optimization framework based on message-passing networks is combined into a unified tracker. TransCenter (Xu et al. 2022) adopts dense representations with image-related dense detection queries and sparse tracking queries.

Our approach inherits the tracking-by-regression paradigm, but we have refrained from using additional graphical optimization or complex motion appearance models. We have designed a temporal coherent object flow that can be integrated into the object detector to form a compact tracking framework.

Method

Our OFTrack, illustrated in Figure 1, consists of three components: a feature extraction backbone, a bounding box regression head (det head), and an object motion prediction head (flow head). We choose the YOLOX (Ge et al. 2021)as our backbone and det head.

Flow head, as illustrated in Figure 2, consists of two parts: **object perception** is designed to improve the semantic awareness of objects. The target sampling generates the object queries of the first frame within the section. The scale-aware correlation calculates the correlation volumes, and the surrounding sampling extracts the surrounding features from correlation volumes centered around target objects; **motion estimation** is proposed for simultaneous motion estimation in multiple frames. It concatenates the object quires, surrounding features and object coordinates for attention layers from spatial and temporal dimensions, respectively. The object motion is updated in multiple iterations.

In addition, we have designed a temporal coherent enhancement to improve the accuracy and robustness of our flow head when simultaneously estimating the motion of multi-frames. **Bidirectional pre-train** is to estimate the trajectories through randomly masking tracks during the pre-training phase. **Curvature constraint** utilizes smoothing constraints of curve to avoid oscillation and uncertainty in motion estimation, making the results more aligned with actual situations.

Motivation

The goal of our flow head is to adapt the optical flow to the object flow, which needs to solve two main challenges: (1) **Object awareness**. Our flow head treats the target as a whole for motion estimation, i.e., all pixels inside the object share the same motion pattern. This requires our flow head to be aware of the object semantics of the target and be adaptive to changes in the target scale; (2) **Long-term association**. The optical flow is designed to estimate the pixel motion at

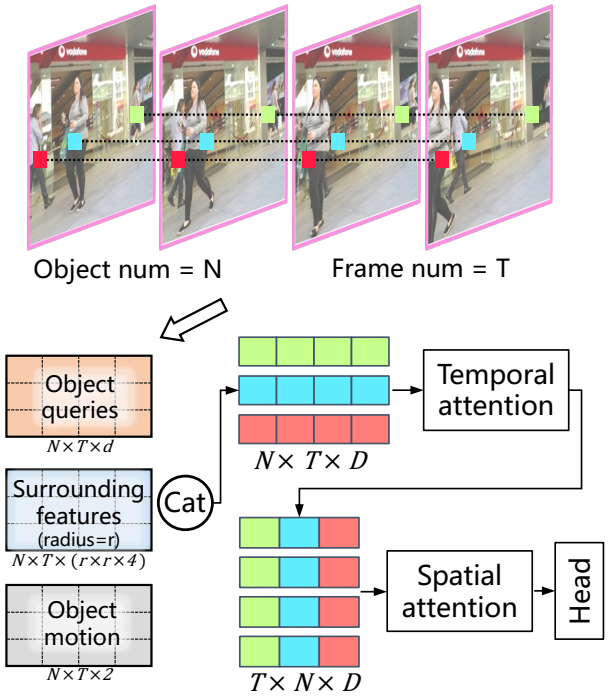


Figure 3: Architecture of motion estimation. Taking a section as the basic process unit, we concatenate of object queries, surrounding features and object motion. Temporal attention models the same object across frames, and spatial attention interacts different objects within one frame.

an infinitesimal distance across a small number of frames, while our object flow in the tracking task requires a motion estimation over an extended period of video sequence.

To address the above challenges, we developed several strategies. **For object awareness**, we attach the flow head to the object detector and jointly train object tracking and detection in an end-to-end manner. This allows the flow head to obtain the object-aware capability from the detector. Additionally, we design the object perception to perform object-aware sampling that focuses more on the semantic information of objects, and multi-scale correlation to enhance the adaptability to object scale changes. **For long-term association**, we propose the spatial-temporal attention in the motion estimation to simultaneously predict the motion of multi-frames across long intervals. To avoid the oscillation and unreliability of long trajectories that are prone to occurring in multi-frame motion estimation, we have designed temporal coherent enhancement during the training process, enabling the model to achieve stronger estimation capabilities and higher long-term trajectory accuracy.

Object Flow Head

Object perception Assuming that we have completed the tracking for the section $k-1$ with the frame length T (frame ids are within the range $[t-T+1, t]$), now our focus is the section k (frame ids are $[t, t+T-1]$). Note that for continuous association, there is an overlap of one frame between

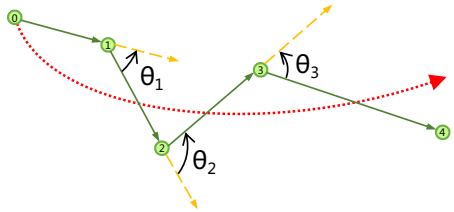


Figure 4: Schematic of curvature constraints. Green line indicates an estimating track, and green points are the observations on it. The red line is the curvature smoothing ground-truth. The yellow lines are the extension of previous trajectory. The goal is to constraint pinch angles θ of each point.

adjacent sections, meaning that the last frame of the previous section serves as the first frame of the current section.

Target sampling processes the feature of frame t, given the frame features $f^t \in \mathbb{R}^{d\times H\times W}$, $f^t \in \mathbb{R}^{d\times H\times W}$, and the tracking result in frame t with the number of objects N, the object center points $C_t = (x_i^t, y_i^t), i \in [1, N]$. We conduct bilinear interpolation from f^t at the locations C_t to get the object query $Q_t \in \mathbb{R}^{d\times N}$.

Scale-aware correlation calculates the correlation volumes between frame t and frames in the section k (take one frame $t_i \in [t, t+T-1]$ for clarity) in two steps: (i) expanding the frame feature f^{t_i} to 4 scales $\{f_1^{t_i}, f_2^{t_i}, f_3^{t_i}, f_4^{t_i}\}$, $f_1^{t_i}$ is the scaled up feature at size $2H \times 2W$ using bilinear interpolation, $f_2^{t_i}$ is in the original feature size f^{t_i}, $f_3^{t_i}$ and $f_4^{t_i}$ are with $H/2 \times W/2$ and $H/4 \times W/4$ resolutions using the average pooling; (ii) calculating the scale-aware correlation volumes between the object query in frame t and the all scaled features in frame t_i, the correlation volumes $\mathbf{V}^{t_i}$ are efficiently computed through the matrix multiplication as:

$$V_k^{t_i} = Q_t^T \cdot f_k^{t_i}, \quad V_k^{t_i} \in \mathbb{R}^{N\times H_k\times W_k}, \quad k \in [1, 4] \quad (1)$$

Surrounding sampling processes the correlation volumes of all frames in the section k (take one frame t_i for clarity), according to the location (x, y) centered on the targets in frame t, we extract the surrounding feature $B_k^{t_i} \in \mathbb{R}^{N\times r\times r}$ on the correlation volumes $V_k^{t_i}$ using bilinear interpolation on the meshgrid G(x,y) with the radius r:

$$G(x, y) = \{(x+dx, y+dy) | \, dx, dy \in \mathbb{Z}, \, dx, dy \le r\} \quad (2)$$

Motion estimation We design the temporal attention and spatial attention for motion estimation, as illustrated in Figure 3. The input tokens are the concatenation of object queries, surrounding features and the object motion, the initial object queries Q_t and object coordinates C_t in frame t are repeated to all frames in the section as $Q \in \mathbb{R}^{N\times T\times d}$ and $C \in \mathbb{R}^{N\times T\times 2}$, the object motion $\eta(C - C_t) \in \mathbb{R}^{N\times T\times d2}$ is obtained by the sinusoidal positional encoding η of coordinates,the surrounding features of radius r extract from the all 4 scaled correlation volumes are $B \in \mathbb{R}^{N\times T\times(r\times r\times 4)}$. The input tokens ($\mathbb{R}^{N\times T\times D}$) are fed into the temporal attention layers first, then reshaped to the $\mathbb{R}^{N\times T\times D}$ to the spatial attention layers, finally the object motion is predicted by a simple linear regression head.

GFC	OP		MOTA	IDF1	HOTA
	TarS	SurS			
✓			71.2	32.9	38.5
	-	-	74.4	55.8	51.9
	✓	-	76.1	68.0	60.7
	✓	✓	78.6	71.7	63.3

(a) Comparisons of object perception. **GFC** means Grid Feature Concatenation, **OP** is our Object Perception, **TarS** is Target Sampling, **SurS** is Surrounding Sampling.

BiP	CurC	MOTA	IDF1	HOTA
-	-	78.6	71.7	63.3
2(8)	-	79.1	74.4	64.3
4(8)	-	79.1	74.7	64.6
6(8)	-	79.8	75.2	65.5
6(8)	✓	80.6	77.4	66.7

(b) Comparisons in temporal coherent enhancement. **BiP** is Bidirectional Pre-train, numbers below are masked tracks (total tracks). **CurC** is Curvature Constraint.

iters	MOTA	IDF1	HOTA
1	72.6	66.8	59.9
2	73.7	68.8	60.5
4	77.5	70.3	62.5
6	78.6	71.7	63.3
8	78.6	71.7	63.3
12	78.8	71.8	63.3

(c) Number of iterations for motion estimation.

Scale strategy	MOTA	IDF1	HOTA
$[1]$	72.6	35.0	40.1
$[1/2, 1]$	73.7	52.1	49.6
$[1/4, 1/2, 1]$	76.1	70.1	62.0
$[1/8, 1/4, 1/2, 1]$	77.8	69.8	62.1
$[1/4, 1/2, 1, 2]$	78.6	71.7	63.3

(d) Multiple scale comparison of scale-aware correlation.

T	S	MOTA	IDF1	HOTA
6	-	71.0	53.2	48.2
-	6	73.3	57.5	54.0
4	4	75.3	68.6	58.3
6	6	78.6	71.7	63.3
10	10	78.6	71.8	63.3

(e) Number of Temporal (T) and Spatial (S) attention layers.

Len	MOTA	IDF1	HOTA
2	76.2	67.5	60.5
4	78.4	71.5	63.2
8	78.6	71.7	63.3
12	76.1	68.4	60.8
16	76.6	66.8	60.0

(f) Sequence length of one section.

Table 1: Ablation experiments. The model is trained on MOT17 train-half and tested on val-half. Default settings are marked in gray. See Section *Ablation Study* for details.

We apply the motion estimation for M times in order to progressively improve the track estimates, that generates a sequence of motion estimation $\{C^1, ..., C^M\}$, the input tokens will be updated by re-sampling the object queries and surrounding features based on the C^i in each iteration. Additionally, we conduct experiments on the effect of iteration number on motion evaluation and select the appropriate iteration number, $N = 6$, considering both accuracy and efficiency under the MOT scenario.

Temporal Coherent Enhancement

Bidirectional pre-train We adopt a two-stage training process, the bidirectional pre-train serves as the first stage to achieve consistent training only for the flow head, and the second stage is to train the whole network (including flow head and det head) in an end-to-end manner.

In the bidirectional pre-train of the flow head, we sample T consecutive frames at random intervals of $[1, 2, 3]$, and randomly select $T_e (T_e < T)$ frames (referred as **masked tracks**) from the input section (the rest are referred as **visible tracks**) for motion estimation. Such a strategy allows for both forward motion estimation in chronological order and reverse estimate in reverse order, increasing the adaptive and generalization capabilities of the flow head. The initial value of the masked tracks are obtained using linear interpolation. Object queries are obtained from the target feature of the first visible tracks. We take the L1 distance between the estimation and ground-truth over the full iterations with exponentially increasing weights for each frame t_i:

$$\mathcal{L}_{pre} = \sum_{i}^{M} \sum_{t_i=0}^{T_e} \gamma^{M-i} \left| C_{t_i}^{i} - C_{t_i}^{gt} \right| \tag{3}$$

where $\{C^1, ..., C^M\}$ represent the all iterations of motion estimation for masked tracks, C^{gt} is the ground-truth and we set $\gamma = 0.8$.

In the second stage, instead of bidirectional estimation, we set the first frame as a reference to predict the subsequent frames in the section, which is consistent with the process of inference. We combine the detector loss $\mathcal{L}_d$ in (Ge et al. 2021) for each frame with motion estimation loss:

$$\mathcal{L} = \lambda_d \sum_{t_i=1}^{T} \mathcal{L}_d^{t_i} + \lambda_f \sum_{i}^{M} \sum_{t_i=1}^{T} \gamma^{M-i} \left| C_{t_i}^{i} - C_{t_i}^{gt} \right| \tag{4}$$

where $\lambda_d = 1.0$ and $\lambda_f = 2.0$ are the weighting factors.

Curvature constraint The curvature constraint is to smooth the motion trajectory, ensuring that the curvature at each successive point as consistent as possible. To align with frame-scenarios where the trajectory consists of discrete points, we calculate the curvature of each point in terms of angle between neighboring paths.

Specifically, considering a set of trajectory points $\{C_0, C_1, C_2, C_3, C_4\}$, as illustrated in Figure 4, the path from C_0 to C_1 is denoted as $\vec{v}_{0\text{-}1}$, the curvature of point C_1 can be described as the angle θ_1 between two neighboring paths $\theta_1 = \langle \vec{v}_{0\text{-}1}, \vec{v}_{1\text{-}2} \rangle$. This pinch angle is expressed within the range $(0, 2\pi)$, allowing us to distinguish cases where the motion deviation on both sides. The curvature smoothing constraint is formulated such that the pinch angles of all points (excluding the first point and last point) on the trajectory tend to be equal, as expressed by the following:

$$\mathcal{L}_{cur} = \sum_{t_i=1}^{T-2} \left| \theta_{t_i+1} - \theta_{t_i} \right| \tag{5}$$

The curvature constraint is used in both two phases of training, so the above loss function in 3 and 4 finally ex-

Methods	Venue	DanceTrack					MOT17					MOT20				
		MOTA	IDF1	HOTA	AssA	DetA	MOTA	IDF1	HOTA	AssA	DetA	MOTA	IDF1	HOTA	AssA	DetA
CenterTrack	ECCV20	86.8	35.7	41.8	22.6	78.1	67.8	64.7	52.2	51.0	53.8	/	/	/	/	/
MOTR	ECCV22	79.7	51.5	54.2	40.2	73.5	71.9	68.4	57.2	55.8	/	/	/	/	/	/
Bytetrack	ECCV22	89.5	52.5	47.3	31.4	71.6	80.3	77.3	63.1	62.0	64.5	77.8	75.2	61.3	59.6	63.4
P3AFormer	ECCV22	/	/	/	/	/	**81.2**	78.1	/	/	/	78.1	76.4	/	/	/
MeMOT	CVPR22	/	/	/	/	/	72.5	69.0	56.9	55.2	/	63.7	66.1	54.1	55.0	/
GTR	CVPR22	84.7	50.3	48.0	31.9	72.5	75.3	71.5	59.1	57.0	61.6	/	/	/	/	/
TransCenter	TPAMI22	86.8	35.7	41.8	22.6	78.1	73.2	62.2	54.5	49.7	60.1	67.7	58.7	/	/	/
OC-SORT	CVPR23	89.4	54.2	55.1	38.0	80.3	78.0	77.5	63.2	63.4	63.2	75.7	76.3	62.4	62.5	62.4
MotionTrack	CVPR23	91.3	53.8	52.9	34.7	80.9	81.1	**80.1**	**65.1**	**65.1**	**65.4**	**78.0**	76.5	62.8	61.8	**64.0**
DiffusionTrack	AAAI24	89.5	47.5	52.4	33.5	82.2	77.9	73.8	60.8	58.8	63.2	72.8	66.3	55.3	51.3	59.9
DiffMOT	CVPR24	**92.8**	63.0	62.3	47.2	**82.5**	79.8	79.3	64.5	64.6	64.7	76.7	74.9	61.7	60.5	63.2
OFTrack	-	90.9	61.7	60.9	45.2	81.9	79.9	77.5	63.1	62.4	63.9	75.3	74.7	61.9	62.1	62.2
OFTrack-ReID	-	91.2	**65.6**	**63.4**	**48.7**	**82.1**	80.1	78.8	64.1	63.3	64.6	75.6	**76.9**	**63.4**	**62.7**	62.9

Table 2: Performance comparison to state-of-the-art approaches on the DanceTrack test set, and the MOT17 and MOT20 test set under the private protocol. The highest-ranking is emphasized in bold.

pressed as:

$$\mathcal{L}_{pre} = \sum_{i}^{M}(\sum_{t_i=0}^{T_e} \gamma^{M-i} \left|C_{t_i}^{i} - C_{t_i}^{gt}\right| + \hat{\alpha}\mathcal{L}_{cur}) \quad (6)$$

$$\mathcal{L} = \lambda_d \sum_{t_i=1}^{T} \mathcal{L}_d^{t_i} + \lambda_f \sum_{i}^{M}(\sum_{t_i=1}^{T} \gamma^{M-i} \left|C_{t_i}^{i} - C_{t_i}^{gt}\right| + \hat{\alpha}\mathcal{L}_{cur}) \quad (7)$$

Considering the distinct characteristics of different datasets, where some datasets exhibit linear trajectories (e.g., MOT17), while others demonstrate nonlinear trajectories (e.g., DanceTrack), we establish a learnable loss weights $\hat{\alpha}$ and set the initial weights to 0.5 for linear trajectories and 0.1 for nonlinear trajectories.

Experiments

In this section, we verify the individual contributions in the ablation study and present the tracking evaluation on several challenging benchmarks, including MOT17 (Milan et al. 2016), MOT20 (Dendorfer et al. 2020), DanceTrack (Sun et al. 2022) and KITTI (Geiger, Lenz, and Urtasun 2012).

Implementation Details

For MOT17 and MOT20 that only consist of pedestrians, we adopt the pretrained YOLOX detector from ByteTrack. For KITTI that are driving scenarios, we adopt the COCO-pretrained YOLOX(Ge et al. 2021) and use the KITTI training set to train the model. DanceTrack is a challenging dataset with highly non-linear motion, and we adopted the same training method as KITTI to train our model. Additionally, we adopt the ReID part (Luo et al. 2019) as the same setting in BoT-SORT (Aharon, Orfaig, and Bobrovsky 2022). The training samples are directly sampled from the same sequence within the interval length of 6. The size of an input image is resized to 1440×800. The flow head parameters are initialized with Xavier Uniform. The AdamW (Loshchilov and Hutter 2018) optimizer is employed with an initial learning rate of 1e-4 and the learning rate decreases according to the cosine function with the final decrease factor of 0.1. We adopt a warm-up learning rate 1e-5 with a 0.2 warm-up factor on the first 5 epochs. We train our model on 4 Nvidia Tesla V100 GPUs for a total of 80 epochs. The mini-batch size is set to 16 with each GPU hosting 4 batches. Our approach is implemented in Python 3.8 with PyTorch 1.10.

Ablation Study

We ablate our approach using the MOT17 dataset. We split the MOT17 train set into train-half set and val-half set as in ByteTrack, all of the ablation experiments are trained on train-half and tested on val-half.

Object perception. We compared the traditional grid feature concatenation (GFC, see Supplementary material for more detail) with our object perception. From Table 1a we find that our object perception has a significant superiority over GFC. Furthermore, adopting the closet grid feature points as the object queries and selecting surrounding features from grid points (line 2) only yields the HOTA of 51.9. The target sampling effectively improves HOTA to 60.7.

Temporal coherent enhancement. We test the bidirectional pre-train and curvature constraint in the Table 1b. We set the section length to 8 and test the different lengths of masked tracks. The the best results is achieved when the length of mask tracks is 6. The curvature constraint can effectively improve the tracking performance.

Iterative motion estimation. We ablate the impact of the different number of iterations for motion estimation. From Table 1c we discover that the performance improvement becomes negligible when the number of iterations exceeds 4, and the impact remains constant after 6 iterations.

Scale-aware correlation. Table 1d shows the results on scale-aware correlation, we up-scale (2x) using the bilinear interpolation and down-scale (1/2x, 1/4x, 1/8x) using the average pooling, all the scaling layers are combined with the original features (1x) for calculation. From Table 1d we can easily find that increasing the number of scales can steadily improve the performance. Notably, with an equal number of

features (4 layers), our object-aware strategy (line 5: HOTA 63.3) outperforms the pooling strategy in the original optical flow (line 4: HOTA 62.1).

	Methods	HOTA	MOTA	DetA	AssA
Car	CenterTrack	73.02	**88.83**	**75.62**	71.20
	TrackMPNN	72.30	87.33	74.69	70.63
	QDTrack	68.45	84.93	72.44	65.49
	QD-3DT	72.77	85.94	74.09	72.19
	Eager	**74.39**	87.82	75.27	74.16
	OFTrack (ours)	73.75	87.73	72.62	**77.71**
Person	MPNTrack	45.26	46.23	43.74	47.28
	CenterTrack	40.35	53.84	44.48	36.93
	TrackMPNN	39.40	52.10	44.24	35.45
	QDTrack	41.12	55.55	44.81	38.10
	QD-3DT	41.08	51.77	44.01	38.82
	Eager	39.38	49.82	40.60	38.72
	OFTrack (ours)	**47.97**	**58.95**	**44.93**	**53.11**

Table 3: Performance comparison to state-of-the-art approaches on the KITTI MOT test set.

Temporal and spatial attention layers. We compare the temporal and spatial attention in Table 1e and observe a poor performance when using them individually. We opt for a final solution comprising 6 layers for both spatial and temporal attention, totaling 12 layers.

Sequence length of one section. Simultaneous multi-frame tracking enhances the predictive ability over long sequences, but excessively long sequences may surpass the precise prediction range. In Table 1f, we can see that the optimal performance are obtained when the length is 8.

State-of-the-Art Comparison

DanceTrack is a long-range group dancing dataset and have frequent crossovers with highly non-linear motion. Table 2 shows that our OFTrack performs superior quality, obtains the 60.9 HOTA score. With a commonly used ReID model, OFTrack-ReID further boost the HOTA to 63.4.

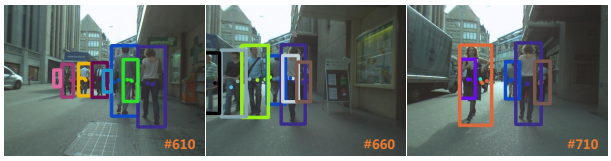
MOT17 and MOT20 are dominant datasets for multi-pedestrian tracking. The performances are presented in Table 2 under the "private" protocol. As can be seen from the comparison, our OFTrack achieves a superior performance both in MOT17 and MOT20 with the HOTA of 63.1 and 61.9, respectively. By incorporating a ReID module, our approach further enhances performance, attaining HOTA scores of 64.1 and 63.4, respectively.

KITTI is a classic tracking benchmark for cars and pedestrians. We show the evaluation in Table 3, In terms of car tracking, we achieve a HOTA score of 73.75. In the pedestrian tracking, our OFTrack demonstrates a notable performance advantage over other trackers, achieving a remarkable HOTA score of 47.97.

Visualization

We offer visualizations of prototypical challenging scenarios, including the non-linear motion scene (Figure 5a), scale-changing scene (Figure 5b), and very crowded scene (Figure 5c) to demonstrate the tracking abilities of the proposed temporal coherent object flow tracker. Figure 5 shows the object boxes of the shown frame (frame number in the lower right corner) and center trajectories of the previous three frames. We observe that our OFTrack has a strong discriminative ability for targets with non-linear motion and keeps high reliable associative ability in crowded scenes.

(a) Non-linear motion scene



(b) Scale change

(c) Crowded scene

Figure 5: Tracking trajectories visualization of Non-linear motion scene in DanceTrack, scale change in MOT17, and crowded scene in MOT20.

Conclusion

In this study, we propose a novel MOT approach OFTrack that exploits an temporal coherent object flow to provide motion estimation. Our OFTrack consists of the object perception and the motion estimation, object perception adopts the object-aware sampling and scale-aware correlation to to enhance the awareness of the object, motion estimation models the correlation of different objects via temporal-spatial attention to achieve robust association in very long videos. Additionally, we design the temporal coherent enhancement strategy, including bidirectional pre-train and the curvature constraint to avoid the oscillation and unreliability during the estimation of long trajectories. The scalability of our object flow allows it to be incorporated into most object detectors for object-level motion estimation. Extensive experiments demonstrate the effectiveness of our flow head.

Limitations. Although our OFTrack can effectively estimate the object motion. We observe that our tracker does not adapt well to the non-rigid deformation of the object. The reason is that our approach treats the target as a rigid object and does not differentiate the information inside it, leading to possible tracking drift when the target aspect ratio varies particularly widely. In future, we intend to integrate our flow head with a deformation estimation counterpart, enhancing its adaptability to a broader range of scenarios.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No.2020YBF2901202), National Natural Science Foundation of China (NSFC No. 62272184 and No. 62402189), the China Postdoctoral Science Foundation under Grant Number GZC20230894, the China Postdoctoral Science Foundation (Certificate Number: 2024M751012), and the Postdoctor Project of Hubei Province under Grant Number 2024HBBHCXB014. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

Aharon, N.; Orfaig, R.; and Bobrovsky, B.-Z. 2022. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*.

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, 3464–3468. IEEE.

Bishop, G.; Welch, G.; et al. 2001. An introduction to the kalman filter. *Proc of SIGGRAPH, Course*, 8(27599-23175): 41.

Brasó, G.; and Leal-Taixé, L. 2020. Learning a Neural Solver for Multiple Object Tracking. In *Proceedings of the CVPR*, 6246–6256.

Cai, J.; Xu, M.; Li, W.; Xiong, Y.; Xia, W.; Tu, Z.; and Soatto, S. 2022. MeMOT: multi-object tracking with memory. In *Proceedings of the CVPR*, 8090–8100.

Cao, J.; Weng, X.; Khirodkar, R.; Pang, J.; and Kitani, K. 2022. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of the ECCV*, 213–229. Springer.

Cetintas, O.; Brasó, G.; and Leal-Taixé, L. 2023. Unifying Short and Long-Term Tracking with Graph Hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22877–22887.

Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; and Leal-Taixé, L. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the CVPR*, 3354–3361. IEEE.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Kumar, R.; Charpiat, G.; and Thonnat, M. 2015. Multiple Object Tracking by Efficient Graph Partitioning. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision*, 445–460.

Liu, Z.; Wang, X.; Wang, C.; Liu, W.; and Bai, X. 2023. SparseTrack: Multi-Object Tracking by Performing Scene Decomposition based on Pseudo-Depth. *arXiv preprint arXiv:2306.05238*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled weight decay regularization. In *Proceedings of the ICLR*.

Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; and Gu, J. 2019. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10): 2597–2609.

Luo, R.; Song, Z.; Ma, L.; Wei, J.; Yang, W.; and Yang, M. 2023. DiffusionTrack: Diffusion Model For Multi-Object Tracking. *arXiv preprint arXiv:2308.09905*.

Lv, W.; Huang, Y.; Zhang, N.; Lin, R.-S.; Han, M.; and Zeng, D. 2024. DiffMOT: A Real-time Diffusion-based Multiple Object Tracker with Non-linear Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19321–19330.

Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; and Feichtenhofer, C. 2022. Trackformer: Multi-object tracking with transformers. In *Proceedings of the CVPR*, 8844–8854.

Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; and Schindler, K. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

Qin, Z.; Zhou, S.; Wang, L.; Duan, J.; Hua, G.; and Tang, W. 2023. MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17939–17948.

Rangesh, A.; Maheshwari, P.; Gebre, M.; Mhatre, S.; Ramezani, V.; and Trivedi, M. M. 2021. Trackmpnn: A message passing graph neural architecture for multi-object tracking. *arXiv preprint arXiv:2101.04206*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Song, Z.; Luo, R.; Yu, J.; Chen, Y.-P. P.; and Yang, W. 2023. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2321–2329.

Song, Z.; Tang, Y.; Luo, R.; Ma, L.; Yu, J.; Chen, Y.-P. P.; and Yang, W. 2024. Autogenic language embedding for coherent point tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2021–2030.

Song, Z.; Yu, J.; Chen, Y.-P. P.; and Yang, W. 2022. Transformer Tracking With Cyclic Shifting Window Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8791–8800.

Sun, P.; Cao, J.; Jiang, Y.; Yuan, Z.; Bai, S.; Kitani, K.; and Luo, P. 2022. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the CVPR*, 20993–21002.

Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; and Luo, P. 2020. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the ECCV*, 402–419. Springer.

Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, 3645–3649. IEEE.

Xu, Y.; Ban, Y.; Delorme, G.; Gan, C.; Rus, D.; and Alameda-Pineda, X. 2022. TransCenter: Transformers with dense representations for multiple-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ye, Y.; Cai, J.; Zhou, H.; Li, G.; Zhang, Y.; Song, Z.; Gao, C.; Yu, J.; and Yang, W. 2024. Progressive Text-to-Image Diffusion with Soft Latent Direction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6693–6701.

Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; and Wei, Y. 2022. Motr: End-to-end multiple-object tracking with transformer. In *Proceedings of the ECCV*, 659–675.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the ECCV*, 1–21. Springer.

Zhang, Y.; Wang, T.; and Zhang, X. 2023. MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22056–22065.

Zhao, Z.; Wu, Z.; Zhuang, Y.; Li, B.; and Jia, J. 2022. Tracking objects as pixel-wise distributions. In *Proceedings of the ECCV*, 76–94. Springer.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3769–3777.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2020. Tracking objects as points. In *Proceedings of the ECCV*, 474–490. Springer.

Zhou, X.; Yin, T.; Koltun, V.; and Krähenbühl, P. 2022. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8771–8780.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.