Multi-Object Tracking with Robust Object Regression and Association

Yi-fan Li[a], Hong-bing Ji[a,**], Xi Chen[b], Yu-kun Lai[c], Yong-liang Yang[b]

[a] *School of Electronic Engineering, Xidian University. Xi'an Shaanxi Province, 710071, China*
[b] *Department of Computer Science, Univeristy of Bath. Bath, BA2 7AY, United Kingdom*
[c] *School of Computer Science and Informatics, Cardiff University. Cardiff, CF24 4AG, United Kingdom*

ABSTRACT

Tracking-by-regression is a new paradigm for online Multi-Object Tracking (MOT). It unifies detection and tracking into a single network by associating targets through regression, significantly reducing the complexity of data association. However, owing to noisy features from nearby occlusions and distractors, the regression is vulnerable and unaware of the inter-object occlusions and intra-class distractors. Thus the regressed bounding boxes can be wrongly suppressed or easily drift. Meanwhile, the commonly used bounding box-based post-processing is unable to remedy false negatives and false assignments caused by regression. To address these challenges, we present to leverage regression tubes as input for the regression-based tracker, which provides spatial-temporal information to enhance the tracking performance. Specially, we propose a novel tube re-localization strategy that obtains robust regressions and recovers missed targets. A tube-based NMS (T-NMS) strategy to manage the regressions at the tube level is also proposed, including a tube IoU (T-IoU) scheme for measuring positional relation and tube re-scoring (T-RS) to evaluate the quality of candidate tubes. Finally, a tube re-assignment strategy is further employed for robust cost measurement and to revise false assignments using motion cues. We evaluate our method on benchmarks, including MOT16, MOT17, and MOT20. The results show that our method can significantly improve the baseline, mitigate the challenges of the regression-based tracker, and achieve very competitive tracking performance.

1. Introduction

Multi-object tracking (MOT) involves localizing objects in each frame and temporally forming trajectories. MOT is one of the core tasks in computer vision to facilitate scene understanding and has various applications such as video surveillance, autonomous vehicles, and human behavior analysis. However, it remains challenging in crowded scenes with occlusions, distractors, low frame rates, and camera motions.

Due to the rapid progress in deep-learning-based object detection, tracking-by-detection became the dominant paradigm in MOT. It divides the MOT task into two separate steps: (i) localizing multiple objects in each frame and (ii) linking the identical objects across frames. Therefore, with the provided detections, MOT is formed as a data association problem. Specifically, it aims to distinguish different targets by assigning a unique identity (ID) while keeping the ID consistent across the temporal domain. Most previous works focus on extracting discriminative features such as appearance features (Wojke et al., 2017; Chen et al., 2018; Zhang et al., 2021), making motion predictions (Zhou et al., 2020; Wu et al., 2021), and utilizing the attention mechanism (Zhu et al., 2018; Guo et al., 2021; Peng et al., 2020a). However, these methods can significantly increase computational complexity since extra tracking-related networks are typically employed.

The tracking-by-regression framework (Bergmann et al., 2019) is proposed to reduce the complexity of data association in MOT. By reusing the regression head of two-stage object detectors Faster R-CNN (Ren et al., 2016), the existing tracks in the previous frame are associated automatically with the targets in the current frame, demonstrating promising tracking performance while significantly reducing the complexity of data association. Later works (Xu et al., 2020; Liu et al., 2020; Stadler and Beyerer, 2021; Guo et al., 2021) take advantage of this new framework for further improvements.

[**] Corresponding author: Tel.: +86-13892821999; fax: (029)88201658; *e-mail:* hbji@xidian.edu.cn (Hong-bing Ji)

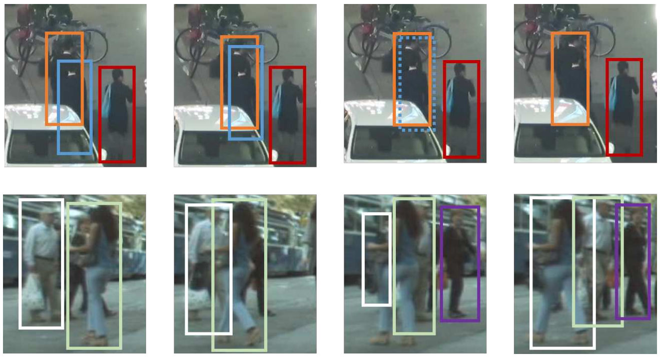


Fig. 1. Noisy features from nearby cause typical false negatives and ID switches in the tracking-by-regression paradigm. The dashed box represents the raw regressed result, which becomes invalid in the post-process. Top row: The nearby targets with similar appearances introduce noisy features. Therefore the dashed blue box drifts to a nearby target and is eventually eliminated. Bottom row: The two boxes in white and green contain mixed features of two targets, and representative features are contaminated as the two targets get closer, resulting in an ID switch. Different box colors indicate different identities. Best viewed in color.

However, despite its simplicity, the regression-based tracking network is modified from an object detector, which relies on the representative features of input targets for regression. As in single object tracking (Li et al., 2018; Yuan et al., 2020) and thermal infrared tracking (Liu et al., 2022), the learned global semantic features are sensitive to all related semantic objects and insensitive to similar objects of the same class. Therefore, the regression-based multi-object tracker is unaware of inter-object occlusions and intra-class distractors, working in an ID-agnostic fashion. That is to say, the noisy features of occlusions and distractors from nearby can easily cause target drift (see Fig. 1) and missing targets, dramatically damaging the tracking performance. Besides, the simple bounding boxes-based post-processing procedure, such as Intersection-over-Union (IoU) and non-maximum suppression (NMS), can easily lead to incorrect relational measurements, ID switches (IDS), false positives (FP), and false negatives (FN) in crowded scenes. Meanwhile, the target identities are generally maintained with NMS based on the confidence obtained from the regression network, which only represents the quality of localization but not tracking and can not sufficiently reveal the quality of identity matching. Therefore, in challenging cases with inter-object occlusions and intra-class distractors, the original post-processing procedure and confidence are not optimal for identity assignment as occluded targets are often suppressed, and identities are often mistakenly assigned. Moreover, the regression-based trackers cannot recover the missed targets once lost.

In this paper, we mitigate these issues in the regression-based trackers and enhance the MOT performance by leveraging regression tubes instead of simple static boxes (see Fig. 2) as input. The regression tubes constructed by consecutive bounding boxes of individual tracks contain spatial-temporal information (such as moving direction and speed) and have been proved to be effective in MOT (Pang et al., 2020) and video object detection (Kang et al., 2016; Tang et al., 2019). We improve the performance of the regression-based tracker by regressing the tubes to unleash the potential of the regression network in three aspects. Firstly, we utilize regression tubes as input and propose a tube re-localization strategy that benefits regression by incorporating spatial-temporal information extracted from the formed tubes. It can effectively resolve the inaccurate regression problem in box-based regression by reusing discriminative features and recovering missed targets. Secondly, we propose a tube-based NMS (T-NMS) mechanism to process the tubes by utilizing the information from the previous frame to process regressed targets at the tube level. With T-NMS, we can promote partly occluded targets to stay active and penalize low-quality regressions, resulting in improved identity consistency. Finally, in order to correctly measure the similarity and assign the identity of targets with regression tubes, we propose a tube re-assignment strategy, which integrates the spatial-temporal features and the motion cues to improve the traditional bounding box-based matching. We evaluate the proposed method through extensive comparisons with existing state-of-the-art trackers and apply our method to other regression-based trackers. Both qualitative and quantitative results demonstrate the effectiveness of our approach.

In summary, the main contributions of this paper are:

- We leverage the regression tubes as input to incorporate the spatial-temporal information and propose a tube re-localization strategy for better target localization and recovering missed targets.
- We propose a T-NMS mechanism that robustly processes the tube regressions at the tube level and correctly scores the associations by penalizing low-quality regressions.
- We propose a tube re-assignment strategy that integrates the spatial-temporal features and motion cues of regression tubes in similarity measuring to revise false assignments to enhance the identity consistency of the tracking.
- Extensive experiments demonstrate that our method effectively improves the performance of the regression-based tracker, and the proposed tracker can obtain very competitive tracking performance on MOT benchmarks.

2. Related Work

The existing MOT algorithms can be categorized into online and offline methods. Future frames can be used for matching globally in offline tracking. Therefore, offline methods are more robust to occlusions and distractors in general. In contrast, only previous and current frames are available for online methods, which have a broader range in real-world applications but are more vulnerable to occlusions and distractors. Here we focus on online tracking where our work lies.

2.1. *Tracking-by-detection*

Most previous methods (Wojke et al., 2017; Chen et al., 2018) utilize the tracking-by-detection (TBD) framework, where off-the-shelf detectors provide the detections on each frame. and then MOT is formulated as a data association problem that links the detected objects with the existing tracks temporarily. A common formalism is to build a graph for associations, where the nodes represent targets, and the edges represent relations of potential links. The appearance features (Wang et al., 2020; Guo et al., 2021; Zhang et al., 2021) are widely used for similarity measuring, and the additional ReID models (Wojke et al., 2017; Son et al., 2017) are often employed to match re-appeared identities to form long trajectories. Besides, Some methods adopt motion models, such as the Kalman filter (Wojke et al., 2017; Zhang et al., 2021), optical flow (Tang et al., 2017), and motion prediction networks (Zhou et al., 2020; Sadeghian et al., 2017; Wang et al., 2022), that incorporate temporal features to make dynamic position predictions to compensate for noisy detections. Some methods establish Recurrent Neural Networks (Milan et al., 2017; Sadeghian et al., 2017; Jain et al., 2020) to model complex motion patterns. Moreover, data association is also formulated as a graph optimization problem in some methods (Li et al., 2022) and solved globally with network flow (Schulter et al., 2017) and Multiple Hypothesis Tracking (Kim et al., 2015) frameworks. Despite the superior performance, this separate pipeline impedes the TBD from real-time utilization.

2.2. *Joint-detection-and-tracking*

The joint-detection-and-tracking (JDT) framework is proposed to eliminate the gap between object detection and data association, reuse backbone features, solve MOT end-to-end with multi-task learning in a single network, and state-of-the-art results are achieved. A unified framework is proposed by (Feichtenhofer et al., 2017) to jointly perform detection and tracking based on the detector R-FCN (Dai et al., 2016). JDE (Wang et al., 2020) employs the detector YOLOv3 (Redmon and Farhadi, 2018) by adding an appearance embedding branch, resulting in detections and representative features of targets obtained with shared backbone features. Likewise, the FairMOT (Zhang et al., 2021) is built on top of the CenterNet (Zhou et al., 2019) with an additional appearance embedding branch. CTracker (Peng et al., 2020a) integrates object detection, feature extraction, and data association in an end-to-end framework. The predictions of the same targets in consecutive frames are obtained with chained boxes. Similarly, CenterTrack (Zhou et al., 2020) is built based on the CenterNet with an added tracking offset prediction branch. DHN (Xu et al., 2020) proposes a Deep Hungarian Net module that transforms the Hungarian algorithm (Kuhn, 1955) to comply with the neural networks by learning to assign the target identities and builds an end-to-end MOT training framework. Although effective, these trackers are all built upon existing detectors by adding tracking-related networks, which brings difficulties in training and increases the parameters of the network. On the contrary, our regression-based tracker with regression tubes as input achieves very competitive results without extra networks and training difficulties.

2.3. *Tracking-by-regression*

This stream works by regressing tracked targets from previous frames for associations by leveraging the regression network of bounding box refinement. The data association is performed without extra matching methods. Tracktor (Bergmann et al., 2019) is an inspiring tracking-by-regression framework that adapts the detector Faster R-CNN (Ren et al., 2016) into a multiple object tracker by re-utilizing the regression head. The existing tracks (represented as bounding boxes) are used as inputs, and the identities of targets are assigned by box regressions directly. This procedure eliminates the necessity of the complex data association process, and some other methods follow this paradigm for further improvements. GSM (Liu et al., 2020) proposes a novel graph representation to leverage the relations among objects to improve the robustness of the similarity model. An occlusion handling strategy that models the relation between occluding and occluded tracks is proposed in TMOH (Stadler and Beyerer, 2021) to improve the track management of the regression-based tracker. The TMOH outperforms the feature-based approaches without a separate re-identification network. TADAM (Guo et al., 2021) utilizes two attention modules that allow the tracker to focus more on targets and suppress the influence of distractors nearby. More discriminative features and accurate position predictions are obtained. However, these methods ignore the spatial-temporal features of tracks and cannot recover the missing targets. Our method leverage the regression tubes as input, fully utilize the spatial-temporal information provided by tubes and retrieve the missing targets to boost the tracking performance.

2.4. *Video object detection and MOT with Tubes*

Video object detection is closely related to MOT without the requirement of identity for each target. However, both tasks face similar challenges, such as occlusions, camera motions, and noisy detections. The tubes incorporating spatial-temporal information and motion cues by stacking consecutive targets are widely used in this field. A tubelet proposal network (Kang et al., 2017) that combines object detection and object tracking is presented to generate tubelet proposals efficiently. The tubes are generated by the object tracker and used for localizing objects. However, our method is different in that tubes are generated from the regression instead of object proposal and are intended for identity assignment. Seq-NMS (Han et al., 2016) is a post-processing strategy that uses high-scoring object detections to boost scores of weaker detections within the same clip, similar to the proposed T-NMS in this paper. However, T-NMS processes the regression tubes based on scores of matching pairs produced by a novel scoring strategy and reveals the quality of candidate tubes. The average detection scores are used in Seq-NMS, indicating the confidence of classification. The tubes are employed in the field of MOT as well. TubeTK (Pang et al., 2020) encodes spatial-temporal features with bounding-tube, regresses bounding tubes for data association, and processes tubes with Tube NMS. However, TubeTK works offline, i.e., the tracking is performed with future frames included for both bounding-tube regression and Tube NMS. An additional IoU-based greedy algorithm is needed to complete data association. However, our method differs because the proposed modules in our method are tailored for regression tubes and track in an online fashion.

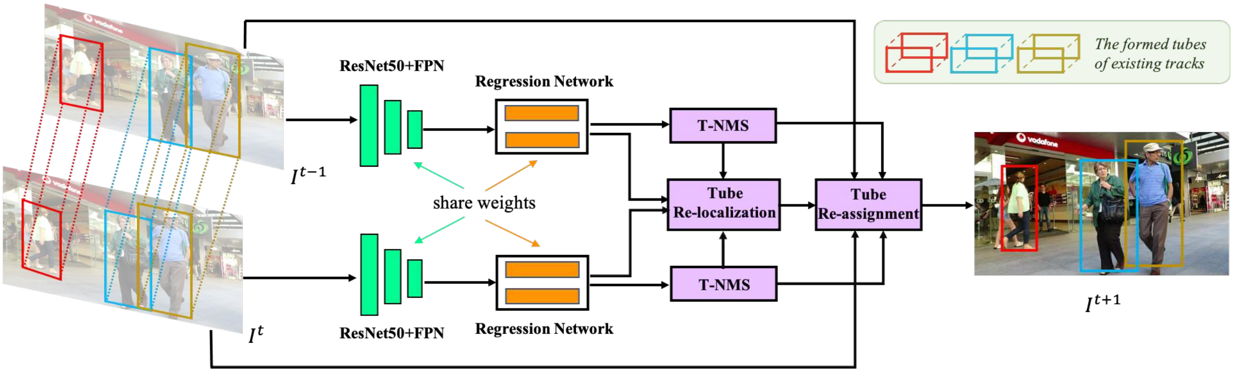


Fig. 2. Overview of the method pipeline. The red, blue, and yellow boxes represent existing tracks with different identities for each frame. The tracked boxes across continuous frames on the frames I^{t-1} and I^t form a tube set (illustrated by boxes connected via dashed lines), which is used as the input for the proposed tube-based robust regression and association. The proposed method extends the tracks to the next frame I^{t+1} with identities. This process is conducted iteratively to the whole sequence. The figure is better viewed in color.

3. Proposed Method

In this section, we first give an overview of our method (Section 3.1), then present its three key components in detail, including tube re-localization (Section 3.2), T-NMS (Section 3.3), and tube re-assignment (Section 3.4).

3.1. Overview

The existing regression-based tracker (Bergmann et al., 2019) performs data association based on the regression procedure of the object detector Faster R-CNN (Ren et al., 2016). Thus heavily depends on the quality of the representative features for the regression head, i.e., the embeddings of existing tracks on previous frames. Specifically, given as input the tracks of the frame I^t represented by a box set $\mathcal{B}^t = \{b_1^t, b_2^t, ..., b_N^t\}$, the regression results of the frame I^{t+1} are $\mathcal{R}_t^{t+1} = \{p_1^t, p_2^t, ..., p_N^t\}$ with identities transferred from the previous frame by regression. We argue that this simple bounding-boxes-based data association is not optimal and cannot robustly handle inter-object occlusions and intra-class distractors since the regression is performed based on the extracted semantic features. Thus it is ID-agnostic and vulnerable. In order to take advantage of spatial-temporal information to tackle the inter-object occlusions and intra-class distractors, we propose to leverage regression *tubes* as the primary input for robust regression and employ general tube-based measuring and processing strategies for robust association in crowded scenes.

The overview of our method is shown in Fig. 2. Compared with previous work (Bergmann et al., 2019) that only utilized the bounding box on the frame I^t, the tracks on frame I^{t-1} are also employed to form the tube set $\mathcal{U}^t$ as the input in our method. Each tube is formed by two boxes with the same identity. The backbones of the siamese network share the weights. To deal with inaccurate regressions and false assignments caused by inter-object occlusions and intra-class distractors, we propose a tube re-localization strategy to process tube-based regressions (Section 3.2), resulting in more accurate regressions and recovers targets of the frame I^{t+1}. We then propose a novel tube-based NMS (T-NMS) (Section 3.3) to better manage the regression tubes at the tube level with the proposed tube IoU (T-IoU) and re-score the corresponding confidences with tube re-scoring (T-RS) scheme. The re-scored confidence is used in tube re-localization (Section 3.2) and tube re-assignment (Section 3.4). The T-IoU is employed when measuring the locality relationships between different tubes. Finally, we propose a tube re-assignment (Section 3.4) strategy to accurately measure the similarity between different targets and revise wrongly assigned identities. The costs of associating candidate tubes are refined by considering spatial-temporal features and motion cues within tubes.

3.2. Tube Re-localization

The regression is performed based on the representative features. Therefore, the regression is an identity-agnostic process and unaware of inter-object occlusion and intra-class distractors. The noisy features from similar appearances and nearby targets can easily damage the regression quality. Consequently, the correct regressed results are often eliminated, and the identities become lost (see the top row in Fig. 1). Thus, one of the drawbacks of the regression-based tracker is that the regression is vulnerable to inter-class occlusions and intra-class distractors. Besides, regression is a one-to-one procedure. Therefore it cannot recover missing targets once lost.

To deal with this, we take regression tubes as input and propose a novel tube re-localization strategy, incorporating spatial-temporal features and historical information of tracked tubes into regression. Better regressions are obtained by reusing the discriminative features inside tubes from previous frames, and missed targets are recovered simultaneously. Technically, given $N(t)$ existing tracks (hereafter referred to as N for simplicity) of the frame I^t denoted as $\mathcal{T}^t = \{T_1^t, T_2^t, ..., T_N^t\}$, each track contains a series of bounding boxes with the same identity. The boxes of N active tracks in frames I^t and I^{t-1} are $\mathcal{B}^t = \{b_1^t, b_2^t, ..., b_N^t\}$ and $\mathcal{B}^{t-1} = \{b_1^{t-1}, b_2^{t-1}, ..., b_N^{t-1}\}$. We utilize N formed tubes $\mathcal{U}^t = \{u_1^t, u_2^t, ..., u_N^t\}$ as the input for regression, where each tube u_i^t is formed by two boxes b_i^{t-1} and b_i^t with the same identity of two consecutive frames. Note that if no track in the previous frame is available, we duplicate the bounding box in the current frame to form the input tube. Based on this, we can regress $\mathcal{U}^t$ and obtain the initial regression results $\tilde{\mathcal{R}}_t^{t+1} = \{\tilde{p}_1^t, \tilde{p}_2^t, ..., \tilde{p}_N^t\}$ and $\tilde{\mathcal{R}}_{t-1}^{t+1} = \{\tilde{p}_1^{t-1}, \tilde{p}_2^{t-1}, ..., \tilde{p}_N^{t-1}\}$ of two frames inside each tube.

To obtain the final regression boxes of each track at the frame I^{t+1}, we need to process the regressed boxes from tubes of each existing track to decide the final regression. The high video frame rate assumption suggests that the regressed boxes with large overlaps and similar geometric characteristics are more likely to be the correct target for each track. Thus, on top of the original regression results, we first coarsely match the regressed boxes in the frame I^{t+1} with the corresponding tracks on frame I^t by minimizing the position cost to obtain the re-arranged regression boxes $\mathcal{R}_t^{t+1}$ with re-assigned identities. A cost matrix C_{pos} can be constructed by measuring the position cost when linking box $\tilde{p}_j^t$ in the frame I^{t+1} to the existing tracked box b_i^t in the frame I^t as:

$$C_{pos}^{i,j} = 1 - \text{IoU}(b_i^t, \tilde{p}_j^t). \tag{1}$$

Then the Hungarian algorithm (Kuhn, 1955) is employed for assignment. The assignment of identities for the regressed boxes in $\tilde{\mathcal{R}}_{t-1}^{t+1}$ to previous existing tracks $\mathcal{B}^{t-1}$ is performed similarly, and we can obtain the re-arranged regression boxes $\mathcal{R}_{t-1}^{t+1}$ as well.

Ideally, it is expected that the two regressed boxes of the same ID in $\mathcal{R}_{t-1}^{t+1}$ and $\mathcal{R}_t^{t+1}$ are largely overlapped with each other with a very similar appearance since each target has only one position on each frame. Thus, we can obtain the final regressions of each tube by measuring the overlaps on the same frame. More specifically, assume there are N formed tubes $\mathcal{U}^t$ at the frame I^t, two regressed box sets with the same assigned ID are $\mathcal{R}_t^{t+1} = \left\{p_1^t, p_2^t, ..., p_N^t\right\}$ and $\mathcal{R}_{t-1}^{t+1} = \left\{p_1^{t-1}, p_2^{t-1}, ..., p_N^{t-1}\right\}$. The corresponding confidence scores $\mathcal{S}_{t-1}^{t+1}$ and $\mathcal{S}_t^{t+1}$, are re-scored with the tube re-scoring (T-RS) strategy (detailed in Section 3.3). Then we calculate the overlaps of boxes that point to the same target at the frame I^{t+1}, and obtain the overlap set $\mathcal{O}^{t+1} = \left\{o_1^{t+1}, o_2^{t+1}, ..., o_N^{t+1}\right\}$, where $o_i^{t+1} = \text{IoU}(p_i^t, p_i^{t-1})$, which is measured by the original bounding box-based IoU. A significant overlap (larger than a pre-defined threshold η_1) of two boxes indicates high confidence in the regressed position. On the contrary, a low overlap (smaller than a pre-defined threshold η_2) means the tube regresses to two different positions. The box with the higher confidence is treated as the final regression of the track, while the ones with lower confidence are likely to be the recovered targets missed by the detector, which will be treated the same as detections from $\mathcal{D}$. By doing so, the final regressed boxes $\mathcal{R}^{t+1} = (r_1^{t+1}, r_2^{t+1}, ..., r_N^{t+1})$ and the retrieved targets $\mathcal{E}^{t+1} = (e_1^{t+1}, e_2^{t+1}, ..., e_L^{t+1})$, and the corresponding scores, $\mathcal{S}^{t+1}$ and $\mathcal{S}_{retr}^{t+1}$ (obtained with the proposed T-RS) can be obtained by merging the final regression results within each tube. The tube re-localization is summarized in Alg. 1. The retrieved targets $\mathcal{E}^{t+1}$ in the frame I^{t+1} will be added to the provided detection set $\mathcal{D}^{t+1}$ for further process.

The regression tubes contain discriminative spatial-temporal features from previous frames, which have been verified effective by the active tracks up to the current. These discriminative and accurate features are reused for regressing the tracked tubes. Thus, the tube re-localization strategy is beneficial in crowd scenes, where nearby targets and backgrounds can severely contaminate the representative features. Moreover, the targets missed by detectors or drift are also recovered. Therefore, we employ the tubes for robust regression-based tracking to eliminate the influence of inter-class occlusions and intra-class distractors. The effectiveness and generalization of tubes are proved in experiments.

```
Algorithm 1 Tube Re-localization
  Input:
    • Tubes $\mathcal{U}^t$ of existing tracks in frame $I^t$
  Output:
    • Regressions $\mathcal{R}^{t+1}$ and corresponding confidences $\mathcal{S}^{t+1}$ in frame $I^{t+1}$
    • Recovered targets $\mathcal{E}^{t+1}$ and corresponding confidence $\mathcal{S}_{retr}^{t+1}$ in frame $I^{t+1}$
Obtain $\tilde{\mathcal{R}}_t^{t+1}, \tilde{\mathcal{R}}_{t-1}^{t+1}$ with tubes $\mathcal{U}^t$ by regression;
Assign identities for $(\tilde{\mathcal{R}}_{t-1}^{t+1}, \mathcal{B}^{t-1})$ and $(\tilde{\mathcal{R}}_t^{t+1}, \mathcal{B}^t)$ according to the position cost in Eqn. 1;
Obtain the $\mathcal{R}_{t-1}^{t+1}$, $\mathcal{R}_t^{t+1}$ for each tube at frame $I^{t+1}$ based on the positional relation within tube;
Obtain the $(\mathcal{R}^{t+1}, \mathcal{S}^{t+1})$ and $(\mathcal{E}^{t+1}, \mathcal{S}_{retr}^{t+1})$ with T-RS based on overlaps $\mathcal{O}^{t+1}$ between different tubes;
Update provided detection set $\mathcal{D}^{t+1}$.
```

3.3. Tube-based NMS

To avoid redundant targets, NMS is often employed as post-processing for identity management (Bergmann et al., 2019; Stadler and Beyerer, 2021; Shuai et al., 2021). Typically, two boxes are treated as pointing to the same target if they are largely overlapped. The correct regression should have higher confidence to survive, which helps to reduce false negatives and ID switches. However, the score of a regressed box mainly represents the confidence of localization, and it cannot well indicate the tracking quality, thus often leading to false identity assignments under frequent inter-object occlusions and intra-class distractors. For example, when two boxes largely overlap, they are regarded as pointing to the same target (but actually not). Since the original NMS only measures the relation of targets at the box level within the same frame, it ignores the spatial-temporal information and historical relations. The correct regressions may be mistakenly suppressed with a lower score. Thus, the original box-based NMS is unsuitable for measuring and matching the regression tubes, which can easily lead to false negatives and ID switches of occluded and intersected targets in crowded scenes.

To address this, we propose the T-NMS (including the tube IoU and tube re-scoring), which is tailored for processing regression tubes, to measure the positional relations between different targets and evaluate the confidence of the candidate tubes at the tube level with spatial-temporal information considered. The proposed tube IoU (T-IoU) coincides with regression tubes and measures the overlaps of different targets by considering the historical positional relations at tube level. More specifically, two regressed boxes in frame I^{t+1} with different IDs have an overlap measured by the original $\text{IoU}(r_i^{t+1}, r_j^{t+1})$, and their corresponding boxes within tubes in the previous frame I^t have an overlap of $o^t_{i,j} = \text{IoU}(b^t_i, b^t_j)$. The proposed T-IoU of the two target tubes can be calculated as follows:

$$o^{t+1}_{i,j} = \begin{cases} \text{IoU}(r^{t+1}_i, r^{t+1}_j) & o^t_{i,j} < \gamma \\ 0.5[\text{IoU}(r^{t+1}_i, r^{t+1}_j) + o^t_{i,j}] & o^t_{i,j} \geq \gamma, \end{cases} \tag{2}$$

where γ is a pre-defined threshold for T-IoU. The T-IOU considers positional relations of previous frames, which reveal the states before current intersections. Thus it enables the tracker to alleviate the influence of the inter-object occlusion by reducing the actual overlaps between occluded targets and relieving those targets from being suppressed by the coarse decision-making process. Therefore, by measuring with the proposed T-IoU instead of the original IoU, the partly occluded targets with lower scores could survive rather than being eliminated with high overlaps, raising the recall and stability of tracking. The examples can be found in Sec. 4.4. Note that the original IoU is used when calculating the final regression within each tube in the proposed tube re-localization since each target only has one unique position at any single frame. And the proposed T-IoU is employed for tube level measurement between different tubes to reduce the false negatives and ID switches.

The proposed tube re-scoring (T-RS) is used to score the confidence of the formed tubes in tube re-localization. This scoring scheme is also used to evaluate the confidence of candidate tubes formed with boxes from regressions $\mathcal{R}^{t+1}$ and existing tracks. In order to accurately reflect the quality of tracking, we introduce Gaussian penalty functions to the original confidence to discourage the difference of potentially associated targets similar to SOT (Yang et al., 2021). We define the relative positional displacement of linked targets from the frame I^t to frame I^{t+1} as follows:

$$\Delta c_t = \frac{\sqrt{(\Delta c^t_x)^2 + (\Delta c^t_y)^2}}{h^t + w^t}, \tag{3}$$

where $\Delta c^t_x = |c^t_x - c^{t+1}_x|$ and $\Delta c^t_y = |c^t_y - c^{t+1}_y|$ are the absolute center displacement of the candidate tubes. Likewise, a relative shape difference of linked tubes can be defined as:

$$\Delta s_t = \frac{\Delta h^t + \Delta w^t}{h^t + w^t}, \tag{4}$$

where $\Delta h^t = |h^t - h^{t+1}|$ and $\Delta w^t = |w^t - w^{t+1}|$ are the absolute height and width differences, respectively. Consequently, the refined scores of the candidate tubes can be computed by:

$$s_{track} = s_{det} \cdot e^{-\frac{\Delta c_t^2}{\sigma_1^2}} \cdot e^{-\frac{\Delta s_t^2}{\sigma_2^2}}, \tag{5}$$

where s_{det} denotes the original confidence obtained from the regression network, the σ_1 and σ_2 are the Gaussian standard deviation. We set $\sigma_1^2 = \sigma_2^2 = 2.0$ experimentally. The T-RS accords with the principle of tracking based on smooth movement and proves to be effective in practice (see Section 4.4).

The proposed T-NMS well matches with tubes and improves evaluation and processing of tracks, enabling the tracker to manage the regression tubes within and among targets at the tube level and obtain the candidate tubes for further assignment. The missed and occluded targets are prevented from being mistakenly eliminated with T-NMS caused by inter-object occlusion. To summarize, T-NMS promotes spatial-temporally consistent tracks with accurate measurement and refined confidence obtained by the proposed T-IoU and T-RS strategies.

3.4. Tube Re-assignment

The target identities are assigned by regression directly in the original regression-based tracker (Bergmann et al., 2019), greatly reducing the complexity of data association. However, the success of this simple process is based on the assumption of high frame rates and constant target velocity. When the assumption breaks (such as low frame rates and large camera motions), inter-object occlusions and intra-class distractors vastly increase, often leading to false assignments, ID switches, and trajectory fragments. Therefore, it would be better if the tracker is aware of false assignments and has the ability to revise them to enhance identity consistency in crowded scenes.

The positional relation of bounding boxes is widely used for measuring the similarity of linked targets in previous methods (Wojke et al., 2017; Wang et al., 2020; Zhang et al., 2021). This simple metric can cope with most targets correctly for easy tracking scenarios. However, owing to the neglect of temporal information and motion cues, this bounding box-based metric is not optimal for handling complex scenes with occlusions and intersections, resulting in false assignments and track fragmentation. The matching cost of identity is calculated by box-based measurement within each tube in Equ. 1. Therefore, the corresponding identities are coarse and unreliable and need to be re-assigned for better identity consistency. The location and scale penalty regarding the size and position changes to re-rank the candidate targets are widely used for smoothing tracks (Yang et al., 2021; Li et al., 2018). Thus, we propose a tube re-assignment strategy to evaluate the similarity of candidate tubes in multi-object tracking scenarios, which can revise false assignments and mitigate the issues of inter-object occlusions and intra-class distractors.

Consider a set of existing tracks $\mathcal{T}^t = \{T^t_1, T^t_2, ..., T^t_n\}$. Each track is composed of a set of bounding boxes $T^t_i = \{b^t_i, b^{t-1}_i, ...\}$, and each box b^t_i is represented by $\{x^t_1, y^t_1, x^t_2, y^t_2\}$, *i.e.*, top-left and bottom-right coordinates. Regression results of tubes and their confidence can then be obtained using the proposed tube re-localization and T-NMS. The candidate tubes for potential links between existing tracks and regressions are formed as input for the tube re-assignment strategy. Each formed candidate pair is also a tube. Therefore, the re-assignment procedure is well matched with tubes and enables the tracker to perform at the tube level with spatial-temporal information.

Unlike most bounding box-based methods that measure locally by overlaps, the tube re-assignment approach takes into account extra information, including the size and displacement of linked boxes within tubes. More specially, given both the i-th existing track with position $b^t_i = (x^t_{i,1}, y^t_{i,1}, x^t_{i,2}, y^t_{i,2})$ in frame I^t and the j-th regression $r^{t+1}_j = (x^{t+1}_{j,1}, y^{t+1}_{j,1}, x^{t+1}_{j,2}, y^{t+1}_{j,2})$ in frame I^{t+1}, the center position, width, and height of the two boxes are $(c^t_{i,x}, c^t_{i,y}, w^t_i, h^t_i)$ and $(c^{t+1}_{j,x}, c^{t+1}_{j,y}, w^{t+1}_j, h^{t+1}_j)$, respectively. Then the size cost of this formed candidate tube is defined as:

$$C_{size}^{i,j} = \frac{|h_i^t - h_j^{t+1}|}{h_i^t} + \frac{|w_i^t - w_j^{t+1}|}{w_i^t}. \tag{6}$$

Likewise, the displacement cost measures the normalized relative displacement of linked center positions as:

$$C_{dis}^{i,j} = \frac{\sqrt{(\Delta c_x^t)^2 + (\Delta c_y^t)^2}}{h_i^t + w_i^t}, \tag{7}$$

where $\Delta c_x^t = |c_{i,x}^t - c_{j,x}^{t+1}|$ and $\Delta c_y^t = |c_{i,y}^t - c_{j,y}^{t+1}|$. The shape cost is the summation of the size cost and the displacement cost:

$$C_{shape}^{i,j} = C_{size}^{i,j} + C_{dis}^{i,j}. \tag{8}$$

The regression tubes contain temporal information and motion trails of tracks, which are vital for distinguishing intersected targets. For challenging cases such as targets becoming occluded while walking, the moving directions of targets are generally different. Therefore, we utilize the motion information in regression tubes and propose a direction cost that measures the difference between the movement of the candidate tubes. Specifically, for the i-th target, we construct a 6-dimension direction vector as $\mathbf{o}_i^t = (c_{i,x}^t, c_{i,y}^t, x_{i,1}^t, y_{i,1}^t, x_{i,2}^t, y_{i,2}^t)$, the elements of $\mathbf{o}_i^t$ denote horizontal, vertical centers, top-left, and bottom-right coordinates. The direction cost between the i-th existing track and the j-th candidate regression can be calculated with the direction vector with formed tubes as:

$$C_{dir}^{i,j} = 1 - \frac{\mathbf{o}_i^{(t-1)\to(t)} \cdot \mathbf{o}_{i,j}^{(t)\to(t+1)}}{||\mathbf{o}_i^{(t-1)\to(t)}|| \cdot ||\mathbf{o}_{i,j}^{(t)\to(t+1)}||}, \tag{9}$$

where $\mathbf{o}_i^{(t-1)\to(t)} = \mathbf{o}_i^t - \mathbf{o}_i^{t-1}$ and $\mathbf{o}_i^{(t)\to(t+1)} = \mathbf{o}_i^{t+1} - \mathbf{o}_i^t$.

Identity consistency can be further enhanced by minimizing the combination of shape cost and the direction cost as $(C_{shape} + \lambda \cdot C_{dir})$, the λ is used to balance two losses. False assignments can be revised and re-assigned by minimizing this combined cost, and the identity consistency is enhanced simultaneously. This proposed fine-grained cost provides very discriminative similarity measurements in distinguishing the occluded and occluding targets, compensating for the weakness of box-based evaluation and mitigating the influence of inter-object occlusions and intra-class distractors. The tube re-assignment strategy is summarized in Alg.2. Note that the proposed tube re-assignment only considers potential associations between the target and its neighbor candidate regressions whose overlap is larger than a pre-defined threshold ξ. Easy cases can be solved successfully with the original regression-based algorithm, and a large re-assigning area will introduce unexpected false assignments and computational burden. The unmatched detections after tube re-assignment are initialized as new tracks.

4. Experiments

In this section, we evaluate our work by extensive ablations and comparisons. We first present the evaluation datasets, metrics, and implementation details. Then we demonstrate the effectiveness of the proposed tube re-localization, T-NMS, and tube re-assignment by qualitative and quantitative ablation studies. Finally, we carefully compare our work with prior works to demonstrate its superior performance and robustness.

Algorithm 2 Tube Re-assignment

```
Input:
  • Existing tracks $\mathcal{T}^t$ in frame $I^t$
  • Regression results $\mathcal{R}^{t+1}$ in frame $I^{t+1}$
  • Provided detections $\mathcal{D}^{t+1}$ in frame $I^{t+1}$
Output:
  • Tracks $\mathcal{T}^{t+1}$ in frame $I^{t+1}$.
Filter the candidate targets to be re-assigned with overlaps higher than $\xi$ from the nearby;
Compute the shape cost $C_{shape}$;
Compute the direction cost $C_{dir}$;
Minimize $(C_{shape} + \lambda \cdot C_{dir})$ to re-assign the IDs.
```

4.1. Datasets and Metrics

We evaluate our work on three benchmark datasets: MOT16, MOT17 (Milan et al., 2016), and MOT20 (Dendorfer et al., 2020) from the MOTChallenge Benchmark. Both MOT16 and MOT17 contain 7 sequences for training with publicly available ground truths and 7 sequences for online testing. However, MOT16 provides frame-wise box detections using DPM (Felzenszwalb et al., 2009), while MOT17 gives more accurate annotations from three detectors: DPM (Felzenszwalb et al., 2009), Faster R-CNN (Ren et al., 2016), and SDP (Yang et al., 2016). The newly released MOT20 contains 4 training and testing sequences, and all the sequences are collected from extremely crowded scenes with frequent occlusions. To make fair comparisons, we conducted all experiments with public detections to avoid the discrepancy introduced by detectors.

The widely used CLEAR MOT Metric (Bernardin and Stiefelhagen, 2008) is adopted for evaluation. Specifically, metrics such as Multi-Object Tracking Accuracy (MOTA), Multi-Object Tracking Precision (MOTP), False Positives (FP), False Negatives (FN), ID switches (IDS), Most Tracked trajectories (MT), Most Lost trajectories (ML), Fragmentation (FM), and ID F1 Scores (IDF1) (Ristani et al., 2016) are assessed. The two most important metrics are MOTA which evaluates tracking coverage, and IDF1, which describes the performance of identity consistency.

4.2. Implementation Details

We take the Tracktor (Bergmann et al., 2019) as the baseline, and our tracker is built on top of the detector Faster R-CNN (Ren et al., 2016) with ResNet-50 (He et al., 2016) and FPN (Lin et al., 2017) as backbones, which are pre-trained on Microsoft COCO (Lin et al., 2014). Then separately fine-tuned on MOT17Det (Milan et al., 2016) and MOT20Det (Dendorfer et al., 2020) datasets, the former is used for evaluation on MOT16/MOT17, and the latter is for MOT20. We only train the network as a general detector, and the training strategy follows the baseline for a fair comparison. Our tracker follows the public protocol and does not initiate new tracks unless provided by the benchmark. All experiments are conducted with RTX 2080 Ti with PyTorch. As for the parameters, in tube re-localization, we decide the regressions with $\eta_1 = 0.9$ and $\eta_2 = 0.5$. In the T-NMS, we set $\gamma = 0.5$ in Eqn. 2. We set $\lambda = 5$ in tube re-assignment to balance two costs and $\xi = 0.6$ for filtering neighboring targets.

4.3. Target Re-localization

To incorporate spatial-temporal and motion information and reuse discriminative features from previous frames for robust regression, we propose to leverage regression tubes as input. In this work, we utilize the shortest tube, i.e., a 2-frame tube across two frames I^{t-1} and I^t, as inputs to the regression network for the frame I^{t+1}, which has been proved to be effective in MOT (Zhou et al., 2020; Peng et al., 2020a) and video object detection (Tang et al., 2019). Intuitively, more discriminative information could be used with longer tubes. However, this may also introduce unexpected noisy features, especially in large camera motion and crowded scenes. Moreover, the computational cost increases with longer tubes as well.

We compare the tracking performances based on tubes with different lengths. The results are shown in Tab. 1. It is obvious that tubes with 2-frame lengths outperform the 1-frame ones in MOTA and IDF1. The single static bounding boxes are used as input in the 1-frame scenario. The performance of the 2-frame tube also demonstrates the superiority of using tubes over static bounding boxes as input for the regression-based tracker, which provides discriminative features and incorporates temporal information and motion cues. Thus, the 2-frame tube can achieve robustness in regression. More targets are retrieved with regression tubes, leading to decreased FN, ML, and increased MT for different detectors. Therefore, the influence of inter-object occlusions and intra-class distractors are alleviated, and the tracking performance is improved. However, with the longer tubes, i.e., 3-frame, no noticeable improvement is observed in MOTA, while the IDF1 drops and IDS increases in all three detectors. We reckon that 3-frame tubes incorporate more redundant and noisy features, increasing false positives and thus damaging identity convergence. The superior performance of 2-frame tubes over 3-frame ones demonstrates that longer tube is not optimal for regression-based tracker since they introduce unexpected false positives and ID switches. Moreover, a longer tube increases the computational burden inevitably. Therefore, we utilize 2-frame tubes in our method.

To demonstrate the effectiveness and generalization of tubes as input for regression with tube re-localization strategy, we apply the 2-frame regression tubes to Tracktor (Bergmann et al., 2019) and DHN (Xu et al., 2020) in both public and private protocols. The private protocol works by employing fine-tuned models as the detector. The variants with regression tubes are denoted as Tracktor† and DHN†. As shown in Tab. 2, significant improvements are achieved in both public and private settings for Tracktor and DHN, especially in the private setting, where better detection results are provided, demonstrating the superiority of regressing tubes for tracking. More significant improvements in private protocol also show that high-quality detection is the key to tracking. Although FP and IDS increased with tubes, a much larger drop in FN and decrease in ML shows that most retrieved targets are true positives and longer trajectories are formed. Tab. 2 also verifies the generalization of tube re-localization in boosting the performance of the regression-based tracking framework.

Table 1. Experiments on different lengths of tubes for regression. The results are obtained on MOT17 training datasets with public detections provided by DPM, Faster R-CNN, and SDP. The arrows here indicate the optimal trend of metrics.

Length	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓
DPM							
1 frame	61.2	62.8	29.5	26.6	2114	41047	375
2 frames	61.7	64.6	29.7	26.0	2187	40403	443
3 frames	61.7	63.4	29.7	25.8	2258	40218	473
Faster R-CNN							
1 frame	63.4	65.8	40.3	17.9	2382	38251	500
2 frames	64.1	67.1	40.8	17.3	2456	37366	519
3 frames	64.1	66.6	40.8	17.2	2475	37372	522
SDP							
1 frame	72.7	69.0	46.2	13.6	2479	27658	523
2 frames	72.8	70.9	46.3	13.5	2573	27341	585
3 frames	72.9	69.9	46.3	13.4	2653	27185	598

Table 2. Experiments on the effectiveness of tubes as input. The results are obtained on the MOT17 training dataset with detections provided by Faster R-CNN. † indicates the variants with the tubes.

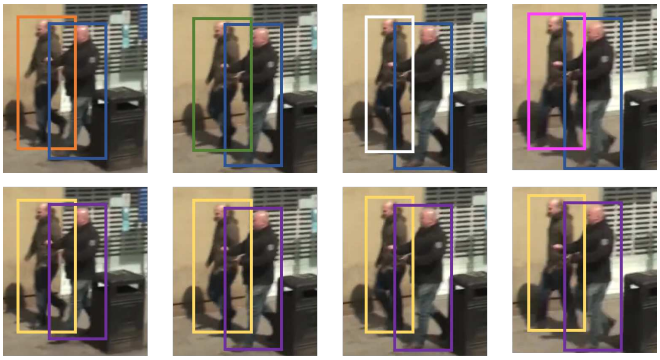
Method	Mode	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓
Tracktor	Public	61.9	64.7	35.3	21.4	323	42454	326
Tracktor†	Public	63.5	66.2	40.5	17.9	2353	38198	484
Tracktor	Privete	70.0	69.6	34.6	8.1	1354	31945	443
Tracktor†	Privete	75.3	72.0	44.3	4.8	5853	21214	702
DHN	Public	62.2	65.7	35.9	14.5	303	41840	272
DHN†	Public	63.9	68.7	43.0	18.0	3325	36749	416
DHN	Privete	70.1	69.2	43.2	7.7	600	32562	400
DHN†	Privete	75.8	74.3	47.4	7.5	6859	19663	630

Since the baseline tracker works in an ID-agnostic fashion due to the unawareness of inter-object occlusions and intra-class differences, the noisy features can damage the regression quality. As shown in Fig. 3 top row, under camera motion, the position of the same target on the adjacent frames differs with low bounding box overlaps, thus introducing the nearby noisy features and leading to inaccurate drift for the left target. As a result, the identities are wrongly assigned, resulting in continuous ID switches. However, with the proposed tube re-localization, as shown in the bottom row of Fig. 3, the regression tubes are used as input, and discriminative features are reused with historical positions considered, resulting in high-quality regressions. Therefore, both targets are tracked correctly and continuously with consistent identities.

We conduct ablation experiments to prove the effectiveness of each component of our method. As shown in Tab. 3, compared with Baseline, the Baseline+TL makes a clear improvement in MOTA and IDF1 with tube re-localization (TL), and the number of FN decreases. Besides, compared with Baseline+TA, the MOTA of Baseline+TA+TL increases by 0.9, the FN decreases dramatically (by 1058), thus more targets are tracked, and longer trajectories are formed with increased MT

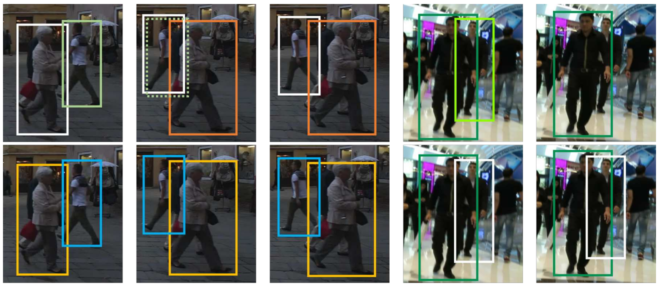
Table 3. Ablation studies of different components of the proposed method. The results are obtained on the MOT17 training dataset with public detectors provided by Faster R-CNN. "TA", "TN", and "TL" stand for the proposed tube re-assignment, T-NMS, and tube re-localization, respectively.

Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓
Baseline	61.9	64.7	35.3	21.4	323	42454	326
Baseline+TA	63.0	66.3	39.7	18.0	2343	38669	482
Baseline+TN	63.4	66.2	40.3	17.9	2352	38241	488
Baseline+TL	63.5	66.2	40.5	17.9	2353	38198	484
Baseline+TN+TA	63.9	66.5	40.5	17.4	2445	37604	547
Baseline+TL+TA	63.9	66.8	40.8	17.4	2434	37611	536
Baseline+TL+TN	64.1	66.7	40.7	17.4	2454	37368	539
Baseline+TL+TN+TA (ours)	64.1	67.1	40.8	17.3	2456	37366	519



(a) 507 (b) 509 (c) 510 (d) 512

Fig. 3. Visualization of the qualitative results with tube re-localization. The tracking results are obtained from frames 507, 509, 510, and 521 of MOT17-13, which are captured with large camera motion. Different box colors represent different identities. Top row: The tracking results without tube re-localization. Bottom row: The results with tube re-localization by regressing tubes to keep identity consistent with camera motion.



(a) 278 (b) 292 (c) 295 (d) 728 (e) 729

Fig. 4. Visualization of the qualitative results with T-NMS (boxes of irrelevant targets are not shown for clarity). The results are from frames 278, 229, and 295 of sequence MOT17-02, and frames 728 and 729 of sequence MOT17-12. Top row: The results without T-NMS. In (a), (b), and (c), the correct regression at frame 292 (dashed green box) is eliminated by NMS. In (d) and (e), the target at frame 729 is suppressed by NMS. Bottom row: The results with T-NMS, where the target survives with the correct identity, and the partly occluded target is preserved.

(by 1.1). The identity convergence is enhanced, proved by the 0.5 increase of IDF1. Similar results can be found by comparing Baseline+TN with Baseline+TL+TN. Likewise, compared with Baseline+TN+TA, the Baseline+TL+TN+TA still improves in MOTA (by 0.2) and IDF1 (by 0.6). Also, lower FN represents that more targets are recovered and tracked, and better trajectories are formed, as shown by increased MT and decreased IDS. Therefore, we argue that the regression tubes are optimal input for the regression-based tracker.

4.4. *Tube-based NMS*

The tube-based NMS (T-NMS) is intended for better processing of regressions at the tube level to improve the robustness of the tracker. T-NMS leverages the historical positions of targets inside the tubes to deal with inter-object occlusions and enhance identity consistency. The T-IoU measures the overlaps at the tube level by considering the positional relations of tubes. For original NMS, if two targets are intersected with substantial overlap, one of the targets would typically be suppressed with a lower score. In contrast, by considering the historical status, we lower the measured overlaps by considering the historical status to make partly occluded targets active to reduce false negatives and ID switches. Besides, the confidences of candidate tubes are vital for identity assignments. The confidence re-scored by T-RS improves data association quality as low-quality regression is suppressed and less likely to survive. Thus false assignments can be largely avoided. Tab. 4 demonstrates the effectiveness of the proposed T-IoU and T-RS. Compared with Baseline, the increased MOTA and decreased FN in the second row demonstrate that T-IoU can keep more true positive targets active, which would be suppressed by the original IoU measurement. Meanwhile, from the third row of Tab. 4, IDF1 is dramatically increased with T-RS, which verifies that T-RS can revise false assignments with re-scored confidence and enhance identity preservation. Further improvements can be achieved with T-IoU and T-RS work together, i.e., the proposed T-NMS. Fig. 4 shows two typical failure cases with false assignments using the original confidence score and original IoU measurement on the top row. The results of utilizing T-NMS are shown in the bottom row. From the bottom row, it is clear that by utilizing T-NMS, the false negatives and ID switches caused by inter-object occlusions and intra-class distractors are resolved, and longer and consistent tracks are formed.

Tab. 3 further demonstrates the effectiveness of the proposed T-NMS. Compared with Baseline, Baseline+TN achieves higher MOTA, IDF1, and MT, as well as lower FN and ML with the proposed T-NMS (TN in table). Besides, Baseline+TL+TN achieves higher MOTA (by 0.6) and IDF1 (by 0.5), lower FN (by 830) compared with Baseline+TL. The proposed T-NMS works parallel with tube re-localization to enhance tracking robustness by keeping more occluded targets alive with correct identities. As a result, longer trajectories with high quality are

Table 4. Ablations on different components of T-NMS. The results on MOT17-02 and MOT17-05. The former sequence is captured with occlusions and target intersections. The latter sequence is recorded with low frame rates, camera motion, and frequent occlusion.

	MOT17-02					MOT17-05				
Method	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓
Baseline	43.9	46.2	62	10288	68	54.9	60.4	228	2713	181
Baseline+T-IoU	44.1	46.3	62	10260	66	54.9	60.6	229	2708	180
Baseline+T-RS	44.0	48.1	60	10288	62	54.9	61.0	248	2707	166
Baseline+T-NMS	44.2	48.5	61	10257	58	55.1	61.4	237	2702	154

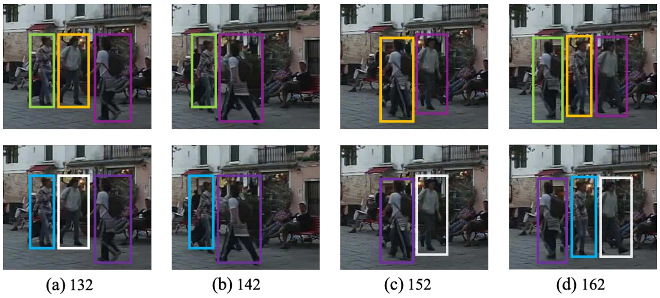

(a) 132 (b) 142 (c) 152 (d) 162

Fig. 5. Visualization of the qualitative results with tube re-assignment. The results are from frames 132, 142, 152, and 162 of sequence MOT17-01. Top row: The results without tube re-assignment make identity association with regression, resulting in ID switches when severe occlusions exist. Bottom row: With tube re-assignment, the motion cues are considered. Thus the identities are assigned correctly.

Table 5. Evaluate the effectiveness of bounding box refinement (BBR), camera motion compensation (CMC), and Re-identification (ReID) for the proposed method and baseline. Here we evaluate the public detections provided by Faster R-CNN on the MOT17 training set.

Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓
Tracktor w/o BBR	57.0	63.1	33.7	22.2	5003	42972	301
Tracktor w/o (CMC+ReID)	61.5	61.1	33.5	20.7	367	42903	1747
Tracktor w/o CMC	61.5	62.8	33.5	20.7	367	42903	921
Tracktor w/o ReID	61.9	64.1	35.3	21.4	323	42454	458
Tracktor	61.9	64.7	35.3	21.4	323	42454	326
Ours w/o BBR	57.3	64.4	39.9	19.8	8211	39423	352
Ours w/o (CMC+ReID)	62.0	62.3	37.9	17.8	2498	38318	1801
Ours w/o CMC	62.6	64.0	37.5	17.8	2502	38321	1200
Ours w/o ReID	63.7	64.7	41.2	17.4	2430	37607	708
Ours	64.1	67.1	40.8	17.3	2456	37366	519

formed with increased MT (by 0.2). Similar enhanced performance can be observed by comparing Baseline+TN+TA with Baseline+TA. Likewise, compared with Baseline+TL+TA, the proposed T-NMS can still help to keep more true positives alive and increase the MOTA (by 0.2), reduce the FN (by 245), improve the IDF1 by 0.3, and IDS reduced at the same time. Therefore, the T-NMS is optimal for processing tubes for robust data association.

4.5. Target Re-assignment

The vulnerable box-based measurement cannot reasonably reflect the relations between tubes, which often leads to ID switches and fragmentations. A typical failure case of ID switches is shown in the top row of Fig. 5. The proposed tube re-assignment is designed by considering association metrics at the tube level within the candidate tube to alleviate the influence of inter-object occlusions and intra-class distractors.

As shown in Tab. 3, the proposed tube re-assignment (TA) can boost the IDF1 when comparing Baseline with Baseline+TA. Furthermore, compared with Baseline+TN and Baseline+TL, both Baseline+TN+TA and Baseline+TL+TA could boost the IDF1 by 0.3 and 0.6, and increase MT by 0.2 and 0.3, respectively. Likewise, compared with the method Baseline+TL+TN, which already achieves good tracking performance in MOTA, the IDF1 of Baseline+TL+TN+TA further increases by 0.4 with the proposed tube re-assignment, and the IDS decreases as well. The superior performance and improvement demonstrate that the proposed tube re-assignment can revise false assignments to re-assign identities correctly, form longer trajectories of high quality, and enhance the identity consistency and stability of the tracker. The bottom row of Fig. 5 shows the case where the mistakenly assigned identities are revised, and consistent trajectories are formed with the proposed tube re-assignment.

4.6. Robustness analysis

The bounding box refinement is utilized in the regression-based trackers (Bergmann et al., 2019; Guo et al., 2021; Stadler and Beyerer, 2021) and other methods (Zhou et al., 2020; Shuai et al., 2021), which aims to refine the noisy detections provided by the benchmark. We evaluate the effectiveness of bounding box refinement for the baseline tracker and our method. As shown in Tab. 5, our method achieves competitive results without bounding box refinement, improving the baseline counterpart tracker in both MOTA and IDF1. Besides, camera motion compensation (CMC) is crucial for compensating camera motion. Compared with the baseline method, our method can better deal with large camera motion and enhance identity consistency with the proposed tube-based regression and tube-level process without CMC. Moreover, the ReID module is an essential component for re-identifying the reappeared targets, which is widely used in the previous state-of-the-art methods (Bergmann et al., 2019; Wojke et al., 2017; Zhang et al., 2020) and proved to be effective in dealing with long-term occlusions and re-appeared targets. Tab. 5 also proves that our method achieves superior results without the ReID module and suppresses the performance of the baseline with the ReID module.

Then we also analyze the robustness of our proposed tracker in crowded scenes that is error-prone by visualizing the comparison with the baseline. We conduct the experiments and visualize examples on the MOT20 test datasets with public detections for a fair comparison. The baseline method Tracktor is

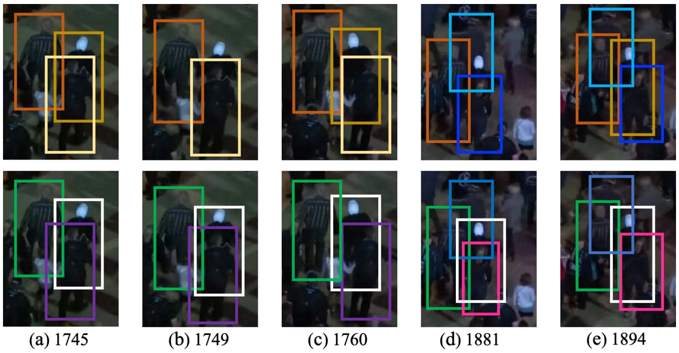


Fig. 6. Qualitative comparison results with the baseline and our method. The tracking results are obtained from frames 1745, 1749, 1760, 1881, and 1894 of test sequence MOT20-04, recorded in a very crowded scene. Different box colors represent different identities. Top row: The tracking results of Tracktor. Bottom row: The results of our method.

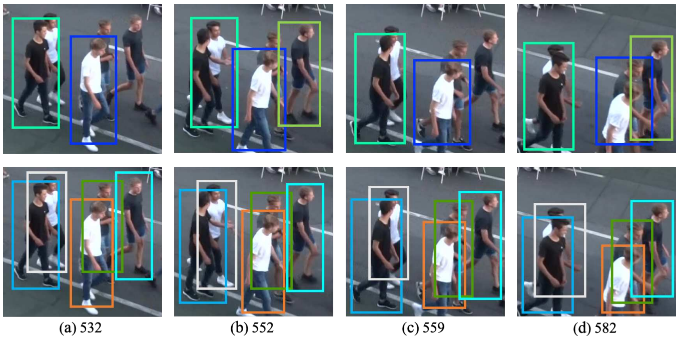


Fig. 7. Qualitative comparison results with the baseline and our method. The tracking results are obtained from frames 532, 552, 559, and 582 of test sequence MOT20-08, which is captured with frequent occlusions. Different box colors represent different identities. Top row: The tracking results of Tracktor. Bottom row: The results of our method.

based on the detector Faster R-CNN with ResNet (Ren et al., 2016) and FPN (Lin et al., 2017) as the backbone network. The Faster R-CNN is aware of the inter-class difference, such as the difference between bicycles and pedestrians, but not the intra-class difference, such as pedestrians with similar appearances from nearby, thus is vulnerable to neighboring distractors of the same class. Besides, it evaluates relations between targets with original NMS at the bounding box level, and the confidence of localization obtained from regression reveals the quality of detection instead of tracking. Therefore the baseline tracker can easily lead to false negatives, false positives, and ID switches. Therefore, we utilize the tubes for regression, and process the targets on the tube level with T-NMS and tube-related metrics.

We visualize extremely crowded cases from MOT20 test sequences obtained from the Tracktor and our method in Fig. 6 and Fig. 7. In the top row of Fig. 6, the target with the white hat is lost in frame 1749, retrieved in frame 1760, and lost again in frame 1881, leading to false negatives. In contrast, this target can be tracked continuously despite being largely occluded in our method. Similar results can be found in Fig. 7. Compared with the results obtained with Tracktor in the top row, our method (bottom row) can significantly reduce the number of FN and ID switches and enhance identity preservation. Therefore, by leveraging the regression tubes, our tracker can mitigate the influence of inter-object occlusions and intra-class distractors, reduce the false negatives and ID switches, recover the lost targets and revise false assignments.

4.7. *Benchmark Comparison*

We extensively evaluate our method by comparing it with the published state-of-the-art (SOTA) methods on multiple benchmark datasets on MOTChallenge Benchmark, including MOT16, MOT17, and MOT20. We adopt the best-performing settings on MOT17 training sets and test on MOT benchmarks with public detections for fair comparisons. We consider only public methods which are comparable to our tracker.

The results are detailed in Tab. 6. Our method achieves very competitive results with public detections. In particular, our method outperforms the baseline Tracktor in terms of MOTA and IDF1 in all three benchmark datasets. The state-of-the-art performance on MOT20 demonstrates the superiority of using regression tubes in dealing with extremely crowded scenes. Compared with methods that are developed from the tracking-by-regression paradigm, including GSM (Liu et al., 2020), DHN (Xu et al., 2020), TADAM (Guo et al., 2021), and TMOH (Stadler and Beyerer, 2021), our method ranks second-best among them that only behind TMOH in MOT16 and MOT17, and achieves the best in MOT20. An occlusion handling strategy that models the relation between occluding and occluded tracks are proposed in TMOH. The inactive tracks are regressed along with the active ones in TMOH. However, our tracker surpasses TMOH in MOT20, showing the advantage of employing tubes for regression in extremely crowded scenes. Besides, as shown in Tab. 7, our method runs faster than TMOH because of fewer computation burdens. Our tracker excels DHN, GSM, and TADAM in MOTA and IDF1, although all of them utilize extra association-related networks. Since our method can retrieve missed targets and keep partly occluded targets active, it tends to have higher FP and IDS. Besides, our method achieves the best performance in terms of FN and ML among them, showing that more true positives are recovered and tracked, and longer trajectories are formed.

Compared with other state-of-the-art methods, ArTIST (Saleh et al., 2021) proposes a stochastic autoregressive motion model that learns the distribution of trajectories, which can inpaints a tracklet in the presence of occlusion and noisy detection. However, our method still excels in IDF1 on MOT17 and MOT20, which shows the effectiveness of our method in enhancing identity consistency. SiamMOT (Shuai et al., 2021) is the current state-of-the-art tracker in MOT17, which integrates the SOT tracker (Li et al., 2018) into Faster R-CNN to form a unified network. The SiamMOT assigns each target a SOT tracker and actively tracks the target once the detector observes. SiamMOT tracks in an ID-aware fashion, thus significantly overcoming the weakness of the detector. In contrast, our method tracks in an ID-agnostic way with limited information provided by the detectors. However, the state-of-the-art performance on MOT20 demonstrates that our method is capable of dealing with highly crowded scenes with frequent occlusions

Table 6. Comparisons with state-of-the-art methods on MOT16, MOT17, and MOT20 datasets with public detections. The ”✓” represents the online method, and the ”✗” denotes the offline method. The best result of each metric is highlighted in bold.

				MOT16				
Method	Mode	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓
STRN (Xu et al., 2019)	✓	48.5	53.9	17.0	34.9	9038	84178	747
Tracktor++ (Bergmann et al., 2019)	✓	54.4	52.5	19.0	36.9	3280	79149	682
DHN (Xu et al., 2020)	✓	54.8	53.4	19.1	37.0	2955	78765	645
Tracktor++v2 (Bergmann et al., 2019)	✓	56.2	54.9	20.7	35.8	**2394**	76844	617
GSM (Liu et al., 2020)	✓	57.0	58.2	22.0	34.5	4332	73573	475
MPN (Brasó and Leal-Taixé, 2020)	✗	58.6	61.7	**27.3**	34.0	4949	70252	**354**
TADAM (Guo et al., 2021)	✓	59.1	59.5	-	-	2540	71542	529
TMOH (Stadler and Beyerer, 2021)	✓	**63.2**	**62.5**	27.0	31.0	3122	63376	635
Ours	✓	62.2	60.9	27.0	**28.1**	5930	**62049**	854
				MOT17				
FAMNet (Chu and Ling, 2019)	✓	52.0	48.7	19.1	33.4	14138	253613	3072
DHN (Xu et al., 2020)	✓	53.7	53.8	19.4	36.6	11731	247447	4792
TPM (Peng et al., 2020b)	✗	54.2	52.6	22.8	37.5	13739	242730	1824
Tracktor++v2 (Bergmann et al., 2019)	✓	56.3	55.1	21.1	35.3	**8866**	235449	1987
GSM (Liu et al., 2020)	✓	56.4	57.8	22.2	34.5	14379	230174	1485
MPN (Brasó and Leal-Taixé, 2020)	✗	58.8	61.7	28.8	33.5	17413	213594	**1185**
TADAM (Guo et al., 2021)	✓	59.7	58.7	-	-	9676	216029	1930
CenterTrack (Zhou et al., 2020)	✓	61.5	59.6	26.4	31.9	14076	200672	2583
TMOH (Stadler and Beyerer, 2021)	✓	62.1	62.8	26.9	31.4	10951	201195	1897
ArTIST (Saleh et al., 2021)	✓	62.3	59.7	29.1	34.0	19611	191207	2062
SimaMOT (Shuai et al., 2021)	✓	**65.9**	**63.3**	**34.6**	**23.9**	18098	**170955**	3040
Ours	✓	61.8	60.4	29.1	27.6	21903	190938	2953
				MOT20				
SORT (Bewley et al., 2016)	✓	42.7	45.1	16.7	26.2	27521	264696	4470
Tracktor++v2 (Bergmann et al., 2019)	✓	52.6	52.7	29.4	26.7	**6930**	236680	1648
ArTIST (Saleh et al., 2021)	✓	53.6	51.0	31.6	28.1	7765	230567	1531
TADAM (Guo et al., 2021)	✓	56.6	51.6	-	-	39407	182520	2690
MPN (Brasó and Leal-Taixé, 2020)	✗	57.6	59.1	38.2	22.5	16953	201384	**1210**
TMOH (Stadler and Beyerer, 2021)	✓	60.1	**61.2**	46.7	17.8	38043	**165899**	2342
Ours	✓	**61.1**	58.9	**48.7**	**17.3**	33108	166170	2192

Table 7. Experiments on the running speed of different methods on the MOT16, MOT17, and MOT20 test sets. The larger the running speed (indicated by Hz), the faster the tracker is.

Dataset	TMOH	TPM	Tracktor	GSM	ArTIST	DHN	Ours
MOT16	0.7	0.8	1.6	7.6	4.5	1.6	1.2
MOT17	0.7	0.8	1.5	8.7	4.5	4.9	1.2
MOT20	0.6	-	1.2	-	1.0	-	0.8

and small targets. The results also prove the generalization ability of our method since the two test scenes of MOT20 never appear in the training set. Thus, we argue that the performance of the regression-based tracker can be considerably improved by regressing and processing tubes.

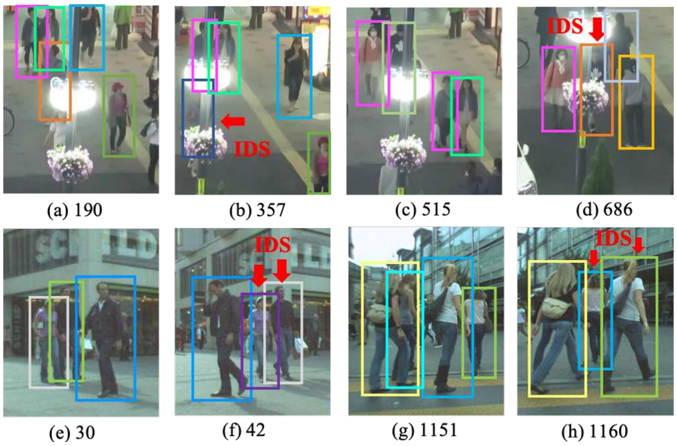
4.8. Discussion

Our method exploits the regression tubes as base inputs for the regression-based tracker. It reuses the discriminative features inside the tubes to eliminate the influence of inter-object occlusions and intra-class distractors and recover missed targets. However, there is still plenty of room for improvement. Fig. 8 shows some typical failure cases of our tracker. The top figures are selected from the test sequence MOT17-03, and the bottom ones are from MOT17-06. For the top row, the lamp strongly influences representative features of targets passing by. Therefore, ID switches are caused in Fig. 8(b) and Fig. 8(d), which are captured from frames 357 and 686. Similarly, in the bottom row, the sequences are captured at a low frame rate, where the positions of the same target in the consecutive frames have comparatively small overlaps, and noisy features from nearby are introduced. The bottom row shows that the reused features are contaminated under continuous occlusion, even using regression tubes as input. As a result, the identities of targets are mistakenly assigned, as shown in Fig. 8(f) and Fig. 8(h). Therefore, the intense illumination and low frame rate are two challenges for our method.

We also compare the running speed of different methods in different MOT benchmarks. As shown in Tab. 7, our method runs slower than baseline since we employ the tubes instead of bounding boxes for regression, more bounding boxes are employed, and extra spatial-temporal information is included for tube processing and similarity calculation. Moreover, our tracker runs faster than TMOH in all three datasets. Our future work is to improve the running speed and efficiency of our method to make it more suitable for real-time utilization.

5. Conclusion

In this work, we proposed to leverage the regression tube as input to address the natural limitations of the tracking-by-regression paradigm for the multi-object tracking. Our method can effectively reuse the discriminative features and spatial-temporal information provided by tubes in dealing with inter-object occlusions and intra-class distractors in crowded scenes. We introduced the tube re-localization strategy, which regressed the tubes of existing tracks to handle inaccurate regressions and recover missed targets. We then presented the T-NMS to mea-



(a) 190 (b) 357 (c) 515 (d) 686

(e) 30 (f) 42 (g) 1151 (h) 1160

Fig. 8. Typical failure cases. Top row: typical failure cases of ID Switches from sequence MOT17-03 caused by illumination changes. Bottom row: typical failure cases of ID Switches from sequence MOT17-06 caused by low frame rate and occlusion. The failure cases are highlighted with red arrows. Different colors represent different identities.

sure and process the tracks at the tube level, which provided an accurate local evaluation between targets, maintained partly occluded targets stay active, and enhanced the consistency of target identities. Benefiting from the improved regressions and re-scored tube confidences, we applied a tube re-assignment strategy that accurately measured the cost of candidate tubes to revise false assignments for robust data association and boosted tracking performance. The results showed that very competitive results are obtained with tubes, which are optimal for the regression-based tracker.

References

Bergmann, P., Meinhardt, T., Leal-Taixe, L., 2019. Tracking without bells and whistles, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 941–951.

Bernardin, K., Stiefelhagen, R., 2008. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing 2008, 1–10.

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking, in: 2016 IEEE International Conference on Image Processing, IEEE. pp. 3464–3468.

Brasó, G., Leal-Taixé, L., 2020. Learning a neural solver for multiple object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6247–6257.

Chen, L., Ai, H., Zhuang, Z., Shang, C., 2018. Real-time multiple people tracking with deeply learned candidate selection and person re-identification, in: 2018 IEEE International Conference on Multimedia and Expo, pp. 1–6.

Chu, P., Ling, H., 2019. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 6172–6181.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, p. 379–387.

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

Feichtenhofer, C., Pinz, A., Zisserman, A., 2017. Detect to track and track to detect, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3038–3046.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2009. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence 32, 1627–1645.

Guo, S., Wang, J., Wang, X., Tao, D., 2021. Online multiple object tracking with cross-task synergy, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8136–8145.

Han, W., Khorrami, P., Paine, T.L., Ramachandran, P., Babaeizadeh, M., Shi, H., Li, J., Yan, S., Huang, T.S., 2016. Seq-nms for video object detection. arXiv preprint arXiv:1602.08465 .

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Jain, M., Subramanyam, A.V., Denman, S., Sridharan, S., Fookes, C., 2020. Lstm guided ensemble correlation filter tracking with appearance model pool. Computer Vision and Image Understanding 195, 102935.

Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., Wang, X., 2017. Object detection in videos with tubelet proposal networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 727–735.

Kang, K., Ouyang, W., Li, H., Wang, X., 2016. Object detection from video tubelets with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 817–825.

Kim, C., Li, F., Ciptadi, A., Rehg, J.M., 2015. Multiple hypothesis tracking revisited, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4696–4704.

Kuhn, H.W., 1955. The hungarian method for the assignment problem. Naval research logistics quarterly 2, 83–97.

Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X., 2018. High performance visual tracking with siamese region proposal network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8971–8980.

Li, W., Wei, Y., Lyu, S., Chang, M.C., 2022. Simultaneous multi-person tracking and activity recognition based on cohesive cluster search. Computer Vision and Image Understanding 214, 103301.

Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

Liu, Q., Chu, Q., Liu, B., Yu, N., 2020. GSM: graph similarity model for multi-object tracking, in: Bessiere, C. (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org. pp. 530–536.

Liu, Q., Yuan, D., Fan, N., Gao, P., Li, X., He, Z., 2022. Learning dual-level deep representation for thermal infrared tracking. IEEE Transactions on Multimedia .

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K., 2016. Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 .

Milan, A., Rezatofighi, S.H., Dick, A., Reid, I., Schindler, K., 2017. Online multi-target tracking using recurrent neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, p. 4225–4232.

Pang, B., Li, Y., Zhang, Y., Li, M., Lu, C., 2020. 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, pp. 6308–6318.

Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y., 2020a. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking, in: European conference on computer vision, Springer. pp. 145–161.

Peng, J., Wang, T., Lin, W., Wang, J., See, J., Wen, S., Ding, E., 2020b. Tpm: Multiple object tracking with tracklet-plane matching. Pattern Recognition 107, 107480.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 .

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence 39, 1137–1149.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking, in: European conference on computer vision, Springer. pp. 17–35.

Sadeghian, A., Alahi, A., Savarese, S., 2017. Tracking the untrackable: Learning to track multiple cues with long-term dependencies, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 300–311.

Saleh, F., Aliakbarian, S., Rezatofighi, H., Salzmann, M., Gould, S., 2021. Probabilistic tracklet scoring and inpainting for multiple object tracking, in:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14329–14339.

Schulter, S., Vernaza, P., Choi, W., Chandraker, M., 2017. Deep network flow for multi-object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6951–6960.

Shuai, B., Berneshawi, A., Li, X., Modolo, D., Tighe, J., 2021. Siammot: Siamese multi-object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12372–12382.

Son, J., Baek, M., Cho, M., Han, B., 2017. Multi-object tracking with quadruplet convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5620–5629.

Stadler, D., Beyerer, J., 2021. Improving multiple pedestrian tracking by track management and occlusion handling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10958–10967.

Tang, P., Wang, C., Wang, X., Liu, W., Zeng, W., Wang, J., 2019. Object detection in videos by high quality object linking. IEEE transactions on pattern analysis and machine intelligence 42, 1272–1278.

Tang, S., Andriluka, M., Andres, B., Schiele, B., 2017. Multiple people tracking by lifted multicut and person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3539–3548.

Wang, H., Li, Z., Li, Y., Nai, K., Wen, M., 2022. Sture: Spatial–temporal mutual representation learning for robust data association in online multi-object tracking. Computer Vision and Image Understanding 220, 103433.

Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S., 2020. Towards real-time multi-object tracking, in: European Conference on Computer Vision, Springer. pp. 107–122.

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric, in: 2017 IEEE international conference on image processing, IEEE. pp. 3645–3649.

Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J., 2021. Track to detect and segment: An online multi-object tracker, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12352–12361.

Xu, J., Cao, Y., Zhang, Z., Hu, H., 2019. Spatial-temporal relation networks for multi-object tracking, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3988–3998.

Xu, Y., Osep, A., Ban, Y., Horaud, R., Leal-Taixe, L., Alameda-Pineda, X., 2020. How to train your deep multi-object tracker, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6787–6796.

Yang, F., Choi, W., Lin, Y., 2016. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2137.

Yang, K., He, Z., Pei, W., Zhou, Z., Li, X., Yuan, D., Zhang, H., 2021. Siamcorners: Siamese corner networks for visual tracking. IEEE Transactions on Multimedia 24, 1956–1967.

Yuan, D., Chang, X., Huang, P.Y., Liu, Q., He, Z., 2020. Self-supervised deep correlation tracking. IEEE Transactions on Image Processing 30, 976–985.

Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W., 2020. Fairmot: On the fairness of detection and re-identification in multiple object tracking. arXiv: 2004.01888 .

Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W., 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision , 1–19.

Zhou, X., Koltun, V., Krähenbühl, P., 2020. Tracking objects as points. European conference on computer vision , 474–490.

Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points. arXiv preprint arXiv:1904.07850 .

Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., Yang, M.H., 2018. Online multi-object tracking with dual matching attention networks, in: European conference on computer vision, pp. 366–382.