Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.056824

ARTICLE

LQTTrack: Multi-Object Tracking by Focusing on Low-Quality Targets Association

Suya Li[1], Ying Cao[1,*], Hengyi Ren[2], Dongsheng Zhu[3] and Xin Xie[1]

[1]Henan Key Laboratory of Big Data Analysis and Processing, School of Computer and Information Engineering, Henan University, Kaifeng, 475001, China

[2]College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing, 210037, China

[3]School of Computer, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

*Corresponding Author: Ying Cao. Email: henu_work_cy@163.com

Received: 31 July 2024 Accepted: 18 September 2024 Published: 15 October 2024

ABSTRACT

Multi-object tracking (MOT) has seen rapid improvements in recent years. However, frequent occlusion remains a significant challenge in MOT, as it can cause targets to become smaller or disappear entirely, resulting in low-quality targets, leading to trajectory interruptions and reduced tracking performance. Different from some existing methods, which discarded the low-quality targets or ignored low-quality target attributes. LQTTrack, with a low-quality association strategy (LQA), is proposed to pay more attention to low-quality targets. In the association scheme of LQTTrack, firstly, multi-scale feature fusion of FPN (MSFF-FPN) is utilized to enrich the feature information and assist in subsequent data association. Secondly, the normalized Wasserstein distance (NWD) is integrated to replace the original Inter over Union (IoU), thus overcoming the limitations of the traditional IoU-based methods that are sensitive to low-quality targets with small sizes and enhancing the robustness of low-quality target tracking. Moreover, the third association stage is proposed to improve the matching between the current frame's low-quality targets and previously interrupted trajectories from earlier frames to reduce the problem of track fragmentation or error tracking, thereby increasing the association success rate and improving overall multi-object tracking performance. Extensive experimental results demonstrate the competitive performance of LQTTrack on benchmark datasets (MOT17, MOT20, and DanceTrack).

KEYWORDS

Low-quality targets association strategy; feature fusion; multi-object tracking; tracking-by-detection

Glossary/Nomenclature/Abbreviations

MOT	Multi-object tracking
TBD	Tracking-by-Detectio
LQA	Low-quality targets association strategy
FPN	Feature Pyramid Networks
MSF F-FPN	Multi-scale feature fusion of Fpn
NWD	Normalized Wasserstein distance

IoU	Intersection over Union
GIoU	Generalized Intersection over Union
LRFS	Labeled Random Finite Sets
Re-ID	Re-identification

1 Introduction

Multi-object tracking (MOT) is a task that forms the tracks of objects by detecting and tracking objects in a video across space and time while maintaining consistent identities [1,2]. It has been utilized in several applications, such as autonomous driving and video surveillance. In real-time related research [3–5], Tracking-by-Detection (TBD) has emerged as one of the mainstream paradigms for target tracking. TBD is a two-stage method involving detection and data association steps. Initially, a detector is employed to identify individual objects in each frame. Subsequently, the detection results are temporally associated using a data association scheme to create continuous tracks for each object. Recently, the rapid advancements in detection and association techniques have led to significant performance improvements in MOT [5–8]. However, occlusion continues to be a significant challenge in MOT, as it can cause objects to become low-quality or even disappear, like the target located by the red boxes shown in Fig. 1. Then the targets with low-quality would cause trajectory interruption and fragmentation, thereby reducing tracking performance. In our paper, the low-quality target is defined by the confidence score of the target where its confidence score lies in $[\tau_{low}, \tau_{high}]$.

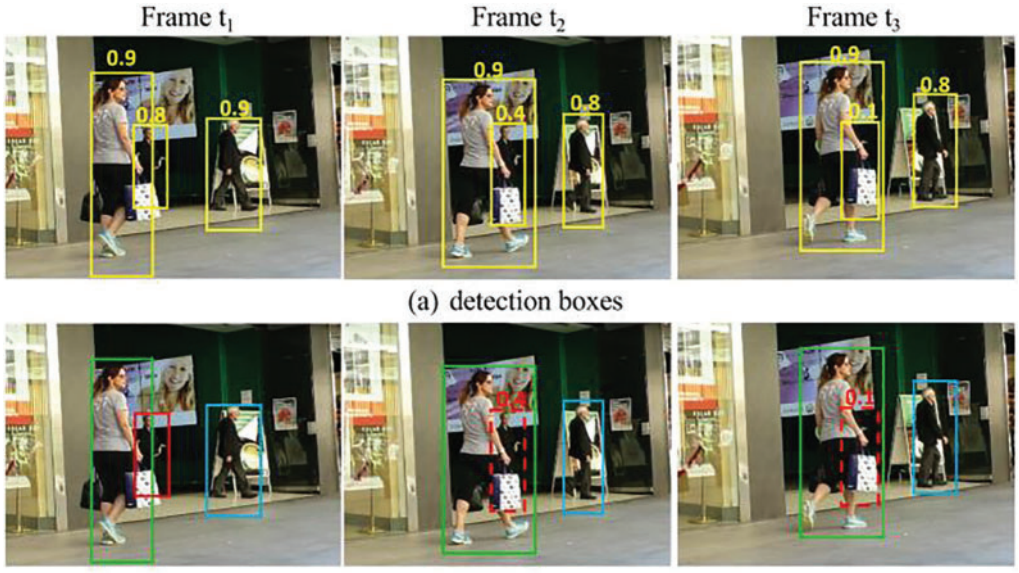

Figure 1: Examples of low-quality targets. (a) shows all the detection boxes with their scores. (b) shows that low score detection is associated because of attention to low-quality targets ($0.1 < \text{score} < 0.6$). Red Dashed box represents associated detection, solid wire frame represents associations. The same box color represents the same identity

Several methods have been proposed to address this. For instance, ByteTrack [5] improves tracking by associating each detection box considering both high and low-score detection. BoT-SORT [9] uses a

simple yet effective method for Intersection over Union (IoU) and Re-identification's (Re-ID) cosine-distance fusion for more robust associations between detections and tracklets. While these recent methods enhance the performance of MOT, issues still remain with the association of low-quality targets caused by occlusion. Firstly, the traditional IoU based measurement is susceptible to positional deviations of low-quality targets. However, occlusion often reduces targets to smaller sizes, leading to a lack of overlap between the bounding boxes of these targets. Consequently, the traditional IoU method fails to accurately reflect the relative similarity between the bounding boxes, resulting in incorrect matching of targets. Secondly, recent methods [7–10] lack the effective consideration for matching the current frame's detection with previously interrupted trajectories from earlier frames $\mathcal{L}_p^{t-\gamma}$, where $\gamma = 2, 3, \ldots, (t-1)$. Algorithms like [11,12] addressed this issue by utilizing multiple hypothesis associations. However, the matching schemes employed in these methods treat all targets similarly and fail to account for attributes of low-quality targets, such as small size. This oversight would limit the further improvement in resolving interrupted trajectories.

Therefore, we construct the LQTTrack with a Low-quality targets association strategy (LQA) to pay more attention to low-quality targets in MOT. In our association design, during the first stage, visual features and motion information of the previous frame's tracklets $\mathcal{L}$ and the current frame's high-quality targets D_{high} are utilized for the initial association. Here, we integrate the multi-scale feature fusion of Feature Pyramid Networks (FPN) [13], named MSFF-FPN, into our model to enrich feature information by aligning semantic features with positional information, thereby improving the success rate of associating high-quality targets with tracklets. During the second association stage, because of the lack of appearance information on low-quality targets with small sizes, motion information is employed to associate the unmatched previous frame's tracklets $\mathcal{L}_u$ and current frame's low-quality targets D_{low}. In this stage, normalized Wasserstein distance (NWD) [14] replaces the traditional IoU to model the bounding box as a two-dimensional Gaussian distribution to measure the motion similarity between $\mathcal{L}_u$ and D_{low}, thus overcoming the limitations of the traditional IoU-based methods that are sensitive to low-quality targets with small sizes and enhancing the robustness of low-quality target tracking. Additionally, benefiting from the thoughts of interrupted trajectories matching, but different with [11,12], the third association stage is proposed to improve the matching between the current frame's low-quality targets, D_{low}, and previously interrupted trajectories from earlier frames $\mathcal{L}_p^{t-\gamma}$, thus can effectively restore the tracklets of low-quality targets, reduce the trajectory interruption phenomena and enhance the overall accuracy of multi-object tracking. Extensive experimental results demonstrate the competitive performance compared to the existing state-of-the-art multi-object tracking methods [5,6,8,15,16] on benchmark datasets (MOT17 [17], MOT20 [18], and DanceTrack [19]). The principal contributions of this work can be summarized as follows:

1) Different from the existing methods, which discard low-quality targets directly, LQTTrack is designed to pay attention to low-quality target association to enhance the overall accuracy of multi-object tracking.

2) In LQA of LQTTrack, besides the employment of multi-scale feature fusion of FPN (MSFF-FPN) in visual features enrichment during the first stage. In subsequent stages, Normalized Wasserstein Distance (NWD) is integrated to address the sensitivity of traditional IoU to low-quality targets with small size and positional deviations, thus improving the robustness of multi-target tracking.

3) In LQA of LQTTrack, the third association stage is integrated to enhance the matching between the current frame's low-quality targets and previously interrupted trajectories from earlier frames to reduce the trajectory interruption phenomena.

2 Related Work

With the continuous advancement of deep learning technology, multi-object tracking techniques have seen rapid improvements in recent years [5,9,20,21]. In the following discussion, we elaborate on two aspects of multi-object tracking: feature extraction, and data association.

2.1 Feature Extraction

Motion information and appearance cues are the main dependent features of current multi-target tracking [22–24]. On the one hand, some researchers opt to forgo appearance information [5,25], relying solely on high-performance detectors and motion information to achieve high operational speed and state-of-the-art performance. For instance, ByteTrack [5] utilizes only motion information, matching tracking through high and low-score detection boxes. Uniform Camera Motion Compensation Track (UCMCTrack) [26] has designed a new motion model-based tracker robust to camera motion, introducing an innovative non-IoU distance metric driven by motion cues alone. On the other hand, numerous researches [6,27] still support the idea that additional appearance cues can enhance multi-object tracking. BoT-SORT [9] proposes camera motion compensation and a more accurate Kalman filter state vector for better bounding box localization, along with a novel fusion method based on IoU and re-id cosine distance. Quasi-Dense Tracking (QDTrack) [28] suggests that position-motion matching is only suitable for simple scenes, as positional information can easily mislead in crowded and occluded scenarios. Choosing to discard position and motion information proposes a matching method based on dense ground truth for extracting appearance features and uses Bi-directional softmax (Bi-softmax) for bidirectional matching, achieving good tracking results using only appearance information. Our method opts to use appearance cues to assist multi-object tracking and embed multi-scale feature fusion to enhance features.

2.2 Data Association

Data association is an important module in the MOT and has also attracted widespread attention and research. Multi-target Tracking using Joint Detection and Tracking (MTTJDT) [29] proposes a multi-loss function that consists of a combination of classifications using the focal loss function and localization loss employing the Complete Intersection Over Union (CIoU) loss function, with a loss-scale parameter used to balance the two functions. This approach enables the prioritization of specific factors, such as object type or position, while also accounting for imbalances in challenging classes and samples. It also employs a dual-regression bounding box to associate objects between adjacent frames by considering the distance between their centers. Transformer-based Assignment Decision Network (TADN) [30] transforms information related to detections and known targets in each frame to directly compute optimal assignments for each detection. Sparse Graph Tracker (SGT) [31] improves tracking of low-score detection by utilizing higher-order relational features, which are more discriminative by aggregating the features of neighboring detection and their relations. ByteTrack [5] makes full use of low-score detection boxes by incorporating them into the process, which improves the accuracy of data association compared to other approaches that only associate high-score detection boxes. However, these methods lack effective consideration for matching the current frame's detection and previously interrupted trajectories, thus limiting the further improvement of tracking performance. To solve this problem, algorithms like [11,12] consider the matching of previously interrupted trajectories through multi-hypothesis data association methods. For instance, Multiple Hypotheses Tracking (MHT) [11] uses a track tree that encapsulates multiple hypotheses starting from a single observation and delays data association decisions by keeping multiple hypotheses active until data association ambiguities are resolved. Tracklet-level Multiple Hypothesis Tracking (TLMHT) [12] incorporates

a tracklet-level association pruning method into MHT and proposes a novel iterative Maximum Weighted Independent Set (MWIS) algorithm to avoid solving the MWIS problem from scratch. Though the effective consideration of interrupted trajectories, the employment of non-differentiate matching schemes ignores the attributes of low-quality targets, and this oversight still forbids further improvement in resolving interrupted trajectories. Therefore, in our paper, the target is firstly divided into high and low-quality ones according to the confidence scores. Then, in our association scheme, the MSFF-FPN is integrated into the first association stage to improve the success rate of high-quality targets. For low-quality targets, the NWD is utilized to assess motion consistency between targets and trajectories, including interrupted ones. This approach ensures that the small size of low-quality targets does not affect the measurement, facilitating the formation of accurate target tracks.

Additionally, to address issues related to data association uncertainty, methods based on Labeled Random Finite Sets (LRFS) model the states and measurements of targets as random finite sets. For instance, Vo et al. [32] introduced the concept of LRFS and proposed a multi-target tracking filter fully described by multi-object prediction and update equations, which is the first theoretical approach capable of trajectory estimation. Additionally, Xue et al. [33] introduced a Bayesian recursive filter tracking method that combines the Density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm with the δ-generalized label multi-Bernoulli (δ-GLMB) filter, which enhances the ability to track clustered targets. Van Ma et al. [34] designed multi-object dynamic and measurement models under the LRFS framework and developed a visual multi-object tracker, based on Generalized label multi-Bernoulli (GLMB) filtering recursion, that can manage track initialization and Re-ID. However, the type of method relies on LRFS to model the target state and measurement information separately, where the transmitted multi-objective probability density iterates over time, resulting in a significant increase in trajectory assumptions, thereby increasing computational complexity. In contrast, compared with the above methods, our data association method employs simple Kalman filtering and similarity calculation to associate and form trajectories. Therefore, it provides enhanced efficiency and adaptability.

Moreover, compared to traditional IoU or distance-based association methods, SimpleTrack [21] adopts Generalized Intersection over Union (GIoU) [35] for association while still utilizing Hungarian or greedy algorithms to match trajectories and detection. However, GIoU degrades to IoU when the predicted and ground truth boxes are completely overlapping, thus failing to capture the relative positional relationship between them. Additionally, GIoU requires the computation of the minimum enclosing rectangle for each predicted and ground truth box, which increases computational complexity and limits convergence speed. In our paper, different from GIoU, NWD is adapted in our paper to measure the motion consistency of trajectories and targets by paying attention to low-quality targets.

3 Method

Different from the existing methods [8,36,37], which discards low-quality targets directly, in this work, LQTTrack with low-quality association strategy (LQA) is constructed to further improve the performance of MOT by paying attention to low-quality targets. The overview of LQTTrack is shown in Fig. 2.

In our LQTTrack, for each frame f_t in video V, a detector (Det) is employed to generate the detection $D_t = \left\{d_j^t\right\}, j = 1, 2, 3, \ldots, n$. Then, the detection boxes are separated into high-quality targets and low-quality targets according to their confidence scores and two score thresholds τ_{low}, τ_{high}, where the score of the high-quality target is higher than τ_{high}, and the score of low-quality targets lies

in $[\tau_{low}, \tau_{high}]$. For each tracklet $l \in \mathcal{L}$, the Kalman Filter (KF) is utilized to predict the position of d_l^t in the current frame for subsequent correlation. Based on these, in our LQA:

First-stage association: In this stage, visual features of the previous frame's tracklets $\mathcal{L}$ and the current frame's high-quality targets D_{high} are employed to measure the association relationship. Considering the varying levels of network architecture, low-level receptive fields are characterized by small spatial dimensions and high resolution, making them adept at handling small-size spatial features. In contrast, high-level receptive fields are large and have lower resolution, which allows them to capture extensive contextual semantic features. Therefore, considering the FPN [13] can integrate the high-level features with low-level features, the thought of multi-scale feature fusion of FPN, named MSFF-FPN, is incorporated into our model to obtain richer appearance information.

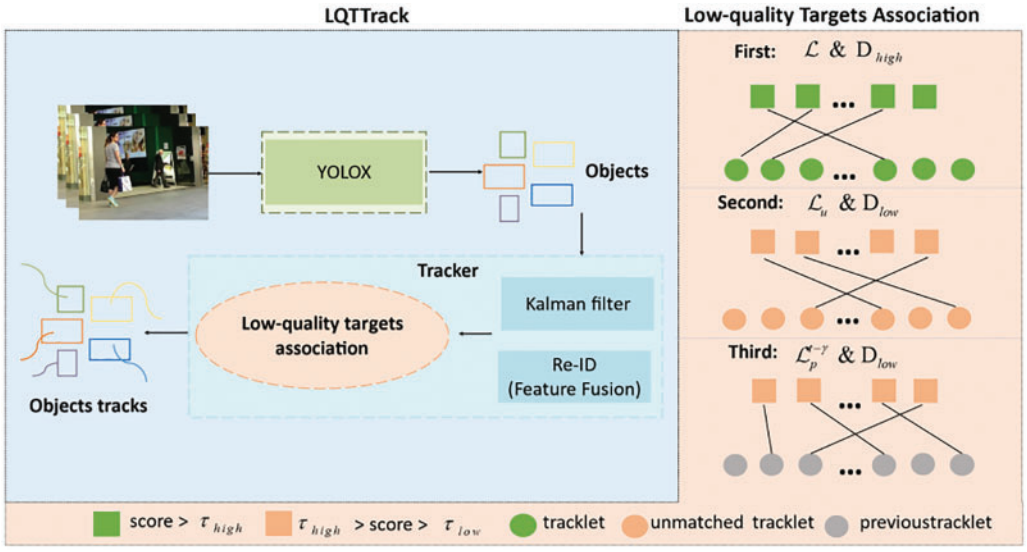

Figure 2: The overview of LQTTrack

Specifically, the multi-scale feature fusion of LQTTrack is shown in Fig. 3. The last three layers of backbone [38] are utilized to output the multi-level visual clues $f_i \in \mathbb{R}^{C\times H\times W}$ $\{i = 1, 2, 3\}$, then, the effective feature representation of detection can be formed by the merge of f_i:

$$f_{[i-1,i]} = Cat\left(upsample\left(Conv\left(f_i\right)\right), Conv\left(f_{i-1}\right)\right) \tag{1}$$

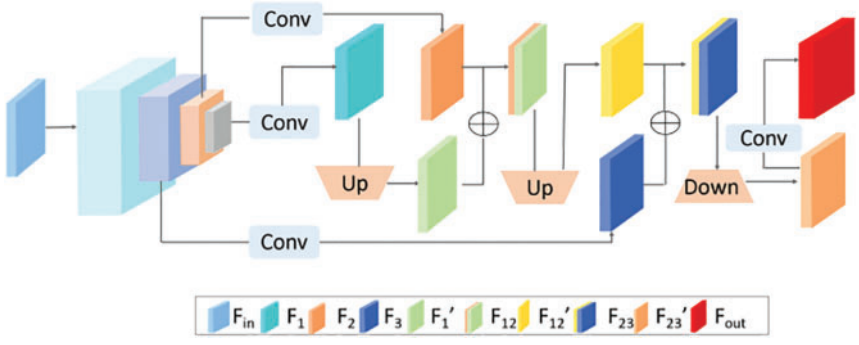

Figure 3: The multi-scale feature fusion of LQTTrack

In Eq. (1), $Conv(\cdot)$ is 1×1 Conv, then, $f_i \in \mathbb{R}^{C \times H \times W}$ is transformed into $\mathbb{R}^{2048 \times H \times W}$, the unsampling operation $upsample(\cdot)$ is used to modify the size of f_i and expand the f_i to twice the original size, the concatenation operation $Conv(\cdot)$ is employed to merge the lower-level feature f_{i-1} with f_i as $[f_{i-1}, f_i]$. Through this way, f_i can be supplied the positional information from the lower-level features to high-level features while retaining the semantic information.

$$f_{final} = Smooth\left(downsample\left(f_{[i-1,i]}\right)\right) \tag{2}$$

Finally, as shown in Eq. (2), downsampling operation $downsample(\cdot)$ and 3×3 convolution operation $Smooth(\cdot)$ are adopted to reduce the noise and redundant information contained in merged features, thereby outputting the final enhanced feature f_{final} for the target with clear information positions and substantial semantic information.

Then, based on the enhanced features, as shown in Eq. (3), the Exponential Moving Average (EMA) is first utilized to form the feature embedding of tracklet $l_i^{t-1} \in \mathcal{L}$.

$$e_{l_i}^{t-1} = \lambda_{t-1} e_{l_i}^{t-2} + (1 - \lambda_{t-1})\, e_{enhanced} \tag{3}$$

In Eq. (3), $e_{l_i}^{t-1}$ represents the appearance embedding of tracklet l_i^{t-1}, $e_{enhanced}$ is the enhanced appearance embedding of the matched detection d_j^{t-1}. λ_{t-1} is a dynamic appearance weighting factor that dynamically adjusts the proportion of visual embeddings between l_i^{t-2} and l_i^{t-1} based on the confidence of different detection boxes, and can be obtained using Eq. (4), in this equation, δ_{det} is the confidence of the detector, σ is the confidence threshold, and λ_f is a fixed value.

$$\lambda_{t-1} = \lambda_f + \left(1 - \lambda_f\right)\left(1 - \frac{\delta_{det} - \sigma}{1 - \sigma}\right) \tag{4}$$

After that, the appearance cost matrix between the current frame high-quality target $d_j^t \in D_{high}, j = 1, 2, 3, \ldots, n$ and previous frames tracklet $l_i^{t-1} \in \mathcal{L}, i = 1, 2, 3, \ldots, m$ is computed based on Eq. (5). In Eq. (5), $e_{d_j}^t$ represents the appearance embedding of the current frame high-quality target $d_j^t \in D_{high}$, $e_{l_i}^{t-1}$ represents the appearance embedding of tracklet l_i^{t-1}, $\bullet$ denotes the dot product, and $\mathbf{C}_a[i,j]$ represents the appearance cost matrix generated by enhanced appearance embeddings, reflecting the appearance similarities between high-quality targets and tracklets, $\mathbf{C}_a[i,j] \in \mathbb{R}^{m \times n}$.

$$\mathbf{C}_a[i,j] = e_{l_i}^{t-1} \bullet e_{d_j}^t \tag{5}$$

Besides the visual similarity, utilising motion information for tracklets prediction through KF, as shown in Eq. (6), motion similarity between prediction box d_l^t of tracklet $l_i^{t-1} \in \mathcal{L}$ and the current frame's detection $d_j^t \in D_{high}$ is computed using the traditional IoU.

$$\mathbf{C}_m[i,j] = e_{l_i}^{t-1} \bullet e_{d_j}^t \tag{6}$$

Then, appearance and motion similarity matrices are integrated according to Eq. (7). Here, a_w is a weighting factor. Subsequently, based on similarity cost matrix $\mathbf{C}[i,j]$, the first matching is conducted using the Hungarian algorithm.

$$\mathbf{C}[i,j] = \mathbf{C}_m[i,j] + a_w \mathbf{C}_a[i,j] \tag{7}$$

Second Stage Association: In this stage, considering the limited visual information of low-quality target $d_j^t \in D_{low}, j = 1, 2, \ldots, k$, the motion information is only used to compute similarity between prediction box d_l^t of unmatched tracklet $l_i^{t-1} \in \mathcal{L}_u$ with low-quality target $d_j^t \in D_{low}$. However, the

traditional IoU metric is highly sensitive to positional deviations, thus when two bounding boxes lack overlap, it would fail to reflect the relative similarity between the two bounding boxes. The low-quality targets with small sizes usually have smaller bounding boxes, making them prone to non-overlapping issues, thereby reducing the effectiveness of IoU matching. To alleviate this issue, we use NWD to measure the similarity between bounding boxes whether the boxes in no-overlap or overlap, thus improving the robustness and accuracy of low-quality target associations.

Specifically, we first model the bounding boxes as two-dimensional gaussian distribution and then use NWD to compute the similarity. Due to the elliptical shape of the density contour of the two-dimensional gaussian distribution, the distribution situation of the gaussian distribution can be represented by the inscribed ellipse of the bounding box. Therefore, the bounding boxes can be modeled as a two-dimensional gaussian distribution, where the center pixel of the bounding box has the highest weight, and the importance of the pixel gradually decreases from the center to the boundary [39]. The bounding box $R = (c_x, c_y, w, h)$ can be modeled into a two-dimensional gaussian distribution $\mathcal{N}(\mu, \Sigma)$ as shown in Eq. (8).

$$\mu = \begin{bmatrix} c_x \\ c_y \end{bmatrix}, \Sigma = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix} \tag{8}$$

Then, the similarity between bounding box A and box B can be converted to the distribution distance between two-dimensional gaussian distributions. According to Wasserstein distance comes from the Optimal Transport theory [14], for two-dimensional gaussian distributions $\mathcal{N}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}_2 = \mathcal{N}(\mu_2, \Sigma_2)$, the 2nd order Wasserstein distance between $\mathcal{N}_1$ and $\mathcal{N}_2$ is defined as in Eq. (9), where $||\cdot||_F$ is the Frobenius norm.

$$W_2^2(\mathcal{N}_1, \mathcal{N}_2) = ||\mu_1 - \mu_2||_2^2 + ||\Sigma_1^{1/2} - \Sigma_2^{1/2}||_F^2 \tag{9}$$

Furthermore, for gaussian distributions $\mathcal{N}_a$ and $\mathcal{N}_b$ which are modeled from bounding boxes $A = (c_{xa}, c_{ya}, w_a, h_a)$ and $B = (c_{xb}, c_{yb}, w_b, h_b)$. We use NWD to measure the similarity between bounding boxes. As shown in Eq. (10), C is a constant closely related to the dataset, $W_2^2(\mathcal{N}_a, \mathcal{N}_b)$ is the 2nd order Wasserstein distance metric.

$$NWD(\mathcal{N}_a, \mathcal{N}_b) = exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{C}\right) \tag{10}$$

Finally, we use NWD to associate prediction box d_l^t of unmatched tracklet $l_i^{t-1} \in \mathcal{L}_u$ and detection box $d_j^t \in D_{low}$, as shown in Eq. (11). Then, based on the cost matrix $\mathbf{C}_u[i,j]$, the matching is conducted similarly to the first stage.

$$\mathbf{C}_u[i,j] = NWD\left(d_l^t, d_j^t\right) \tag{11}$$

Third Stage Association: For the current frame's low-quality target $d_j^t \in D_{low}, j = 1, 2, \ldots, n'$, $n' < k$, different from the related methods, the third association stage is integrated to enhance the matching between the current frame's low-quality target $d_j^t \in D_{low}, j = 1, 2, \ldots, m', m' < k$ and prediction box d_l^t of previously interrupted trajectory $l_i^{t-\gamma} \in \mathcal{L}_p^{t-\gamma}, \gamma = 2, 3, \ldots, (t-1)$ from earlier frames, where $\mathcal{L}_p^{t-\gamma} \in \mathcal{L}_u$, thus reducing the trajectory interruption phenomena and enhancing the overall accuracy of multi-object tracking. As shown in Eq. (12), where d_l^t exists time interval below threshold γ from d_j^t.

$$\mathbf{C}_p[i,j] = NWD\left(d_i^t, d_j^t\right) \tag{12}$$

Then, based on the cost matrix $\mathbf{C}_p[i,j]$, according to the threshold ε to update the matched tracklets $\mathcal{L}$. Through this way to mitigate the issue of trajectory fragmentation caused by occlusions.

The pseudo-code is demonstrated in Algorithm 1, in the designed algorithm, for the input video sequence V, along with an object detector Det and Kalman Filter KF, two thresholds τ_{high}, τ_{low}, ψ, ε are set for different stages of an association scheme. The output of the algorithm is the tracks $\mathcal{L}$ of the video. Firstly, we predict detection boxes & scores, and then, Kalman Filter is used to predict new locations of tracklets of $l_i^{t-1} \in \mathcal{L}$ (Lines 1 to 18). Then, the first stage association (Lines 19 to 28) is performed to match the tracklet $l_i^{t-1} \in \mathcal{L}$ with high-quality targets $d_j^t \in D_{high}$; after that, the second stage association (Lines 29 to 36) is carried out to align the unmatched tracklet $l_i^{t-1} \in \mathcal{L}_u$ and low-quality targets $d_j^t \in D_{low}$. Next, the third stage association (Lines 37 to 43) is additionally executed to add the consideration of the match between the interrupted tracklet from the earliest frames $l_i^{t-\gamma} \in \mathcal{L}_p^{t-\gamma}$ and low-quality targets $d_j^t \in D_{low}$. Finally, the unmatched detection is initialized as new tracklets and removed tracklets that exceed the max age from the tracking list (line 44).

Algorithm 1: Low-quality targets association strategy

Input: Video Sequences V, Detection results Det, Kalman Filter KF, Thresholds τ_{high},τ_{low},ψ,ε
Output: Tracks $\mathcal{L}$

1. Initialization: $\mathcal{L}$
2. **for** *frame* f_t *in* V **do**
3. /* *predict detection boxes & scores* */
4. $D_t \leftarrow Det(f_t)$
5. $D_{high} \leftarrow \phi$
6. $D_{low} \leftarrow \phi$
7. **for** d_j^t *in* D_t **do**
8. **if** $d_j^t.score > \tau_{high}$ **then**
9. $D_{high} \leftarrow \cup\{d_j^t\}$
10. **end**
11. **else if** $d_j^t.score > \tau_{low}$ **then**
12. $D_{low} \leftarrow \cup\{d_j^t\}$
13. **end**
14. **end**
15. /* *predict new locations of tracks* */
16. **for** l_i^{t-1} *in* $\mathcal{L}$ **do**
17. $d_l^t \leftarrow KF(l_i^{t-1})$
18. **end**
19. /* *First stage association: Associate tracklets and high-quality targets by IoU & appearance* */
20. **for** d_j^t *in* D_{high} *and* l_i^{t-1} *in* $\mathcal{L}$ **do**
21. *obtained* $\mathbf{C}[i,j]$ *of* d_l^t *and* d_j^t *according to Eq. (7)*
22. **if** $\mathbf{C}[i,j] > \psi$ **then**
23. $\mathcal{L} \leftarrow \cup\left\{l_i^{t-1}\right\} \leftarrow \cup\left\{d_j^t\right\}$

(Continued)

Algorithm 1 (continued)

```
            end
            else
                    L_u ← ∪{l_i^{t-1}}
            end
        end
        /*Second stage association: Associate unmatched tracklets L_u and low-quality targets
        D_low by NWD*/
        for d_j^t in D_low and l_i^{t-1} in L_u do
            obtained C_u[i,j] of d_l^t and d_j^t according to Eq. (11)
            if C_u[i,j] > ε then
                    L ← ∪{l_i^{t-1}} ← ∪{d_j^t}
                    delete: l_i^{t-1} from L_u
            end
        end
        /* Third stage association: Associate current frame detection with an interval of γ frames
        L_p^{t-γ} ∈ L_u and low-quality targets D_low by NWD */
        for d_j^t in D_low and l_i^{t-γ} in L_p^{t-γ} do
            obtained C_p[i,j] of d_l^t and d_j^t according to Eq. (12)
            if C_p[i,j] > ε then
                    L ← ∪{l_i^{t-γ}} ← ∪{d_j^t}
            end
        end
        Initialized unmatched detection as new tracklets and clear the unmatched tracklets
    end
    Return: L
```

4 Experiments

4.1 Setting

4.1.1 Datasets

We conducted a fair evaluation of several publicly available datasets, including MOT17 [17], MOT20 [18], and DanceTrack [19]. MOT17 and MOT20 are both pedestrian tracking datasets with predominantly linear motion. Notably, MOT20 has a significantly higher density of pedestrians, and the crowded scene means more occlusion, making it a challenging dataset to track. The main task of DanceTrack is to track actors on stage with complex patterns of target movement and large amplitude of movement while multiple targets are dressed in the same costume with similar appearances. For ablation studies, we follow by using the first half of each video in the training set of MOT17 for training and the last half for validation.

4.1.2 Compared Algorithms

In this section, ByteTrack [5], TLMHT [12], GLMB [34], TADN [30], SGT [31], StrongSORT++ [6], Observation-Centric SORT (OC-SORT) [16], FairMOT [8], RelationTrack [40], Correlation Tracker (CorrTracker) [41], and Transformer for MOT (TransMOT) [42] are compared with the proposed LQTTrack. Among these, TADN [30] introduces a transformer-based assignment detection

network as an alternative to traditional data association methods for MOT. SGT [31] and ByteTrack [5] focus on improving the tracking of low-score detections. TLMHT [12] addresses trajectory interruption issues through multiple hypotheses data association. GLMB [34] develops a multi-target tracker using the LRFS framework. Based on this analysis, the experimental section specifically analyzed and compared with ByteTrack, TLMHT, GLMB, TADN, and SGT.

4.1.3 Metrics

This experiment employs CLEAR metrics [43], including Multiple object tracking accuracy (MOTA), Higher order tracking accuracy (HOTA) [44], Identification F1 (IDF1), Identity switches (IDSW) [45], etc., to evaluate the tracking performance comprehensively in various aspects. MOTA emphasizes the tracker's performance, while IDF1 measures the tracker's ability to maintain consistent IDs (Identity documents). We also emphasize the use of Association accuracy score (AssA) to evaluate the association performance. On the other hand, HOTA achieves a balance between detection accuracy, association accuracy, and localization accuracy, making it an increasingly important metric for evaluating trackers. False positives (FP) represents the number of false positives in the entire video sequence, while False negatives (FN) represents the number of false negatives. IDSW denotes the number of identity swaps among tracked targets. Additionally, Frames processed per second (FPS) is utilized to evaluate our tracker's speed (FPS).

4.1.4 Implementation Details

Inspired by ByteTrack's high and low score matching framework and considering the effective use of appearance features, we revised Deep OC-SORT using ByteTrack's matching strategy. We used the revised version as the baseline. To ensure a fair comparison of tracking performance, we employed the same yolox detector as in recent works [5,6,9]. For Re-ID, we used fast-reid [46] with the SBS-50 model, trained with its default training strategy on MOT17, MOT20, and DanceTrack for 60 epochs. For experiments on MOT17 and MOT20, we set a_w to 1.25 for adaptive weighting and 2.25 for DanceTrack. The low detection score threshold τ_{low} was set to 0.1 for MOT17, MOT20, and DanceTrack. The high detection score threshold τ_{high} is set to 0.6 for MOT17 and DanceTrack, and 0.4 for MOT20. Across all experiments, λ_f is fixed as 0.95 for a dynamic appearance, threshold ψ is set to 0.3 for the first association, and threshold ε is 0.6 for the second and third associations to accommodate varying confidences of low-quality targets.

4.2 Benchmark Evaluation

In this section, we present benchmark results for multiple datasets. We conduct experiments on MOT17 [17], MOT20 [18], and DanceTrack [19]. The best results for each indicator are displayed in bold. $\uparrow / \downarrow$ respectively indicate that higher/lower is better. The baseline represents the revised version of Deep OC-SORT after using ByteTrack's high and low score matching strategy, which is the baseline of the LQTTrack.

4.2.1 MOT17

In this part, MOT17 is first used to verify the performance of LQTTrack. For the experiments on the MOT17-test, we employed a proprietary detector to generate detection and aligned them with ByteTrack [5] to ensure fairness. Experimental results and comparisons of MOT17 are shown in Table 1. Through the analysis, it is easy to find that LQTTrack achieves the best performance in 64.8, 80.6, 65.8, and 1008 in HOTA, IDF1, AssA, and IDSW. Compared to TLMHT [12], which addresses

tracklet interruption but does not distinguish between high and low-quality targets, LQTTrack places a greater emphasis on low-quality targets. Our method improves upon TLMHT by 28.7, 24.1, and 399 in MOTA, IDF1, and IDSW, respectively. These improvements further demonstrate the superior performance of our approach relative to TLMHT. Compared to GLMB [34], which uses the LRFS framework, our method surpasses GLMB by 5.9, 5.4, 9.1, and 2247 in HOTA, MOTA, IDF1, and IDSW. Compared to TADN [30], which also addresses occlusion issues, our method surpasses TADN by 24.7, 31.6, and 3861 in MOTA, IDF1, and IDSW. Compared to SGT [31] and ByteTrack [5], which also focus on low-quality targets, our method exceeds SGT by 2.9, 7.8, and 3098 in MOTA, IDF1, and IDSW. Additionally, our method outperforms ByteTrack by 1.7, 3.3, 3.8, and 1188 in HOTA, IDF1, AssA, and IDSW. Moreover, compared to the baseline, LQTTrack improved the method by 1.1, 1.7, and 2.2 in HOTA, IDF1, and AssA. The superior performance demonstrates that the proposed LQTTrack has the ability to improve the association success rate for low-quality targets and optimize the accuracy of MOT by leveraging a more reasonable low-quality targets association strategy.

Table 1: Comparison with state-of-the-art MOT methods on the MOT17 test set

	MOT17							
Tracker	HOTA↑	MOTA↑	IDF1↑	FP(10^4)↓	FN(10^4) ↓	IDSW↓	ASSA↑	FPS↑
FairMOT [8]	59.3	73.7	72.3	2.75	11.70	3303	58.0	25.9
RelationTrack [40]	61.0	73.8	74.7	2.80	11.86	1374	61.5	8.5
CorrTracker [41]	60.7	76.5	73.6	2.98	9.95	3369	58.9	15.6
TransMOT [42]	61.7	76.7	75.1	3.62	9.32	2346	59.9	9.6
StrongSORT++ [6]	64.4	79.6	79.5	2.79	8.62	1194	64.4	7.1
OC-SORT [16]	63.2	78.0	77.5	**1.51**	10.70	1950	63.2	**29.0**
TLMHT [12]	–	50.6	56.5	2.22	25.50	1407	–	–
GLMB [34]	58.9	73.9	71.5	2.51	11.90	3255	–	–
TADN [30]	–	54.6	49.0	3.63	21.49	4869	–	–
SGT [31]	–	76.4	72.8	2.60	10.29	4101	–	–
ByteTrack [5]	63.1	**80.3**	77.3	2.55	**8.37**	2196	62.0	29.6
Baseline	63.7	79.3	78.9	1.68	9.92	1074	63.6	11.3
LQTTrack(ours)	**64.8**	79.3	**80.6**	1.60	9.97	**1008**	**65.8**	11.8

4.2.2 MOT20

In this section, MOT20 is employed to further evaluate the performance of the proposed. Unlike MOT17, MOT20 contains more crowded scenes where higher occlusion implies more low-quality targets and higher chances of unmatched targets. The experimental results presented in Table 2, LQTTrack still surpasses the current SOTA (State-of-the-art) algorithms and achieves 64.0, 78.9, and 65.8 in HOTA, IDF1, and AssA. Compared with ByteTrack [5], our method outperforms 2.7, 3.7, 6.2, and 386 in HOTA, IDF1, AssA, and IDSW. Additionally, compared to SGT [31], our method outperforms 2.7, 8.3, and 1637 in MOTA, IDF1, and IDSW. Compared to GLMB [34], our method surpasses GLMB by 9.8, 7.8, 11.6, and 2074 in HOTA, MOTA, IDF1, and IDSW. Moreover, compared to the baseline, LQTTrack improved the method by 2.6, 4.4, 4.0, 3.3, and 973 in HOTA, MOTA,

IDF1, AssA, and IDSW. When facing scenarios with many low-quality targets, our processing of low-quality targets significantly improves the accuracy of multi-target tracking. All of these indicate the effectiveness of the proposal.

Table 2: Comparison with state-of-the-art MOT methods on the MOT20 test set

	MOT20							
Tracker	HOTA↑	MOTA↑	IDF1↑	FP(10^4)↓	FN(10^4) ↓	IDSW↓	ASSA↑	FPS↑
FairMOT [8]	54.6	61.8	67.3	10.34	8.90	5243	54.7	13.2
RelationTrack [40]	56.5	67.2	70.5	6.11	10.46	4243	56.4	2.7
CorrTracker [41]	–	65.2	69.1	7.94	9.59	5183	–	8.5
TransMOT [42]	61.9	77.5	75.2	3.42	**8.08**	1615	60.1	–
StrongSORT++ [6]	62.6	73.8	77.0	**1.66**	11.79	**770**	64.0	1.4
OC-SORT [16]	62.4	75.7	76.3	1.91	10.59	942	62.0	**18.7**
SGT [31]	–	72.8	70.6	2.52	11.30	2474	–	–
GLMB [34]	54.2	67.7	67.3	2.96	13.45	2911	–	–
ByteTrack [5]	61.3	**77.8**	75.2	2.62	8.76	1223	59.6	17.5
Baseline	61.4	71.1	74.9	4.78	10.02	1810	62.5	2.3
LQTTrack(ours)	**64.0**	75.5	**78.9**	1.74	10.87	837	**65.8**	1.7

4.2.3 DanceTrack

In this part, DanceTrack is utilized to validate the performance of the proposed. DanceTrack is a dataset with the challenges of complex object motion patterns and similar appearances. The validated results and comparisons are exhibited in Table 3. From this presentation, we can see that the proposed achieves better results of 92.0 and 82.1 in MOTA and Detection accuracy (DetA) when compared with the majority of trackers. The better performance demonstrates the effectiveness of our proposed method. However, LQTTrack did not show a significant improvement in HOTA and AssA compared to the baseline. The analysis indicates that using the same parameters (specifically, a_w and τ_{high}) as in MOT17 led to an increase in missed detection, resulting in comparable or slightly lower metrics such as HOTA. This finding demonstrates that the proposed method needs further optimization for better generalization in complex motion scenes.

Table 3: Comparison with state-of-the-art MOT methods on the DanceTrack test set

	DanceTrack			
Tracker	HOTA↑	MOTA↑	DetA↑	AssA↑
FairMOT [8]	39.7	82.2	66.7	23.8
CenterTrack [47]	41.8	86.8	78.1	22.6
QDTrack [28]	45.7	83.0	72.1	29.2
OC-SORT [16]	55.1	92.0	80.3	38.3

(Continued)

Table 3 (continued)

	DanceTrack			
Tracker	HOTA↑	MOTA↑	DetA↑	AssA↑
ByteTrack [5]	47.3	89.5	71.6	31.4
Baseline	**59.6**	90.9	81.5	**43.7**
LQTTrack(ours)	58.9	**92.0**	**82.1**	42.3

Furthermore, with the IDF1-MOTA-HOTA comparisons displayed in Fig. 4, the superior performance, especially in MOT20, indicates the robust performance in handling numerous low-quality targets.

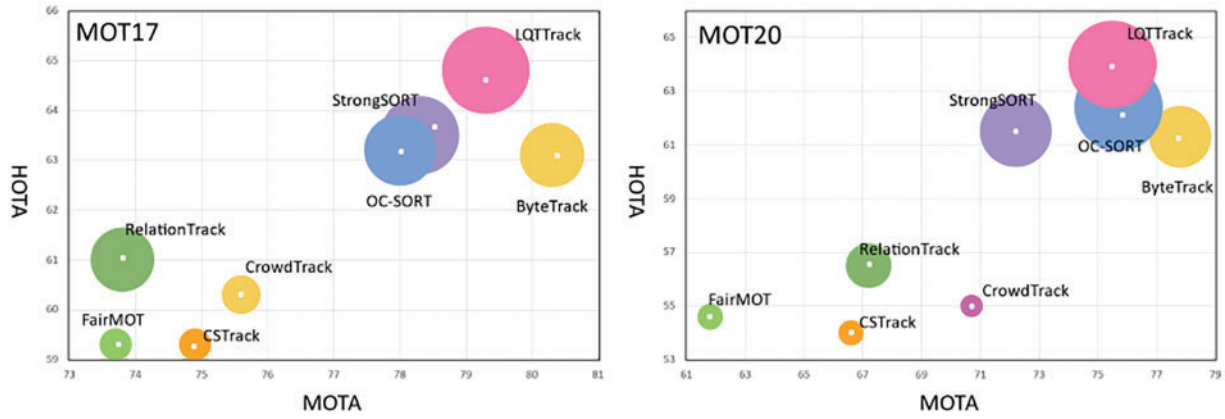

Figure 4: IDF1-MOTA-HOTA comparisons of state-of-the-art trackers with our proposed LQTTrack on the MOT17 and MOT20 test sets. The x-axis is MOTA, the y-axis is HOTA, and the radius of the circle is AssA. There is still excellent performance on MOT20 test sets with many low-quality targets

4.3 Benchmark Evaluation

In this section, ablation studies are conducted to verify the contributions of the designed Low-quality targets association strategy (LQA), Multi-Scale Feature Fusion of FPN (MSFF-FPN) and Normalized Wasserstein distance (NWD) which are the main modules in LQTTrack. Additionally, to avoid potential detector bias, we uniformly employed bytetrack's yolox-x ablation study weights, which were trained on the first half of sequences from CrowdHuman and MOT17.

(1) The effect of LQA.

i) LQA with NWD without MSFF-FPN. We used LQA as a universal tool into the Baseline, Deep OC-SORT, and OC-SORT methods to verify the effectiveness of LQA. The validation results based on the MOT17 dataset are shown in Table 4. Compared to the baseline, the using of LQA improved the method baseline by 1.39, 0.45, 0.54, and 0.95 in HOTA, MOTA, IDF1, and AssA. Compared to Deep OC-SORT, the method of Deep OC-SORT was improved by 0.2, 1.91, 0.57, and 1.51 in HOTA, MOTA, IDF1, and AssA. Similarly, compared to OC-SORT, using LQA improved the OC-SORT method by 0.83, 0.65, 1.64, 1.81, and 68 in HOTA, MOTA, IDF1, AssA, and IDSW. The above performance improvements indicate that LQA is superior in handling low-quality targets.

Table 4: Ablation study on MOT17-val

					MOT17				
Method	original	MSFF	NWD	LQA	HOTA↑	MOTA↑	IDF1↑	AssA↑	IDSW↓
Baseline	✓				68.97	78.37	82.73	73.06	104
	✓	✓			70.29	78.36	83.22	73.95	89
	✓		✓	✓	70.36	78.82	83.27	74.01	99
	✓	✓	✓	✓	70.61	78.79	83.66	74.40	84
Deep OC-SORT	✓				70.20	76.93	82.78	72.54	95
	✓	✓			70.51	78.85	83.41	74.10	81
	✓		✓	✓	70.40	78.84	83.35	74.05	95
	✓	✓	✓	✓	70.56	78.79	83.49	74.28	83
BoT-SORT	✓				68.88	78.40	81.50	71.20	177
	✓	✓			69.34	78.50	82.40	72.00	159
OC-SORT	✓				65.95	75.06	76.78	67.07	320
	✓		✓	✓	66.78	75.71	78.42	68.88	252

Furthermore, video MOT20-03 is selected to further verify the effectiveness of our module on low-quality targets for its dense scenes and many low-quality targets. In this part, we set the detection of $\tau_{low} <$ score $< \tau_{high}$ as low-quality targets, as shown in the Fig. 5, after the second stage which integrates NWD, the low-quality targets are below 612 compared with the baseline. And after the third stage which additionally considers the current frame's unmatched targets and tracklets of previous early frames, it is obvious to find that the number decreases again, and the low-quality targets decrease by 1282 after the third stage. All of these depict the effectiveness of the proposed in optimizing the accuracy of MOT.

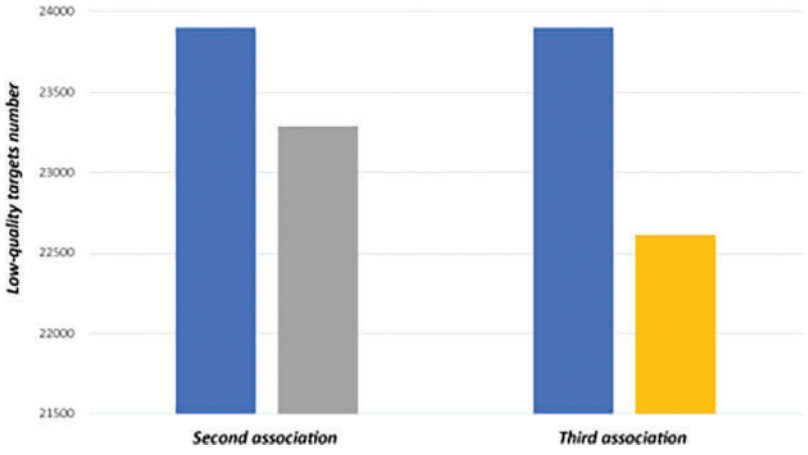

Figure 5: Example of the advantage of LQTTrack. In the second and third associations, we processed low-quality targets separately. LQTTrack effectively reduces the number of low-quality targets on MOT20-03

ii) LQA without NWD and MSFF-FPN. Furthermore, we embed LQA without NWD into the Deep OC-SORT and OC-SORT methods to verify the effect of LQA without NWD. The validation results based on the MOT17 dataset are shown in Table 5. Compared to Deep OC-SORT, the utilizing of LQA improves the method Deep OC-SORT by 0.14, 2.15, 0.39, and 1.19 in HOTA, MOTA, IDF1, and AssA. Similarly, compared to OC-SORT, the method OC-SORT has been improved by 0.6, 1.53, 2.23, and 109 in HOTA, IDF1, AssA, and IDSW. The above performance improvements indicate that LQA without NWD and MSFF-FPN can significantly improve the original performance, which validates the effectiveness of the proposed third-stage association.

Table 5: Ablation study on MOT17 with LQA*. LQA* represent LQA without NWD and MSFF-FPN

	MOT17				
Tracker	HOTA↑	MOTA↑	IDF1↑	AssA↑	IDSW↓
Deep OC-SORT	70.20	76.93	82.78	72.54	95
Deep OC-SORT(LQA*)	70.34	79.08	83.17	73.73	90
OC-SORT	65.95	75.06	76.78	67.07	320
OC-SORT(LQA*)	66.55	74.63	78.31	69.30	211

iii) The effect of LQA with both NWD and MSFF-FPN. Moreover, the proposed association is completely integrated into Baseline and Deep OC-SORT. From the results described in Table 4 for Baseline and Deep OC-SORT, the integration of the proposed has improved the performance of baseline by 1.64, 0.42, 0.93, 1.34, and 20 in HOTA, MOTA, IDF1, AssA, and IDSW, the same optimization of Deep OC-SORT by 0.36, 1.86, 0.71, and 1.74 in HOTA, MOTA, IDF1, and AssA. This further validates the effectiveness and superiority of LQTTrack.

The optimization performance of LQA, both with and without NWD and MSFF-FPN, demonstrates LQTracker's superiority in reducing track fragmentation and tracking errors.

(2) The effect of MSFF-FPN. This work utilizes the multi-scale feature fusion strategy of FPN (MSFF-FPN) to achieve rich feature information. Therefore, MSFF-FPN is embedded in Baseline, Deep OC-SORT, and BOT-SORT. The experimental results are shown in Table 4. Compared to the baseline, the using of MSFF-FPN improves the method baseline on MOT17 datasets by 1.32, 0.49, and 0.89 in HOTA, IDF1, and AssA. Compared to the Deep OC-SORT method, the method of Deep OC-SORT on MOT17 datasets has been improved by 0.31, 1.92, 0.63, and 1.56 in HOTA, MOTA, IDF1, and AssA. Similarly, compared to the Deep OC-SORT, using MSFF-FPN improves the BoT-SORT method on MOT17 by 0.46, 0.9, and 0.8 in HOTA, IDF1, and AssA. These results demonstrate the effectiveness of MSFF-FPN.

(3) The effect of NWD. To address the incomplete consideration of low-quality targets in IoU measurement methods, NWD is adopted to handle low-quality targets with small sizes. Therefore, to measure the contribution of NWD, this section integrates GIoU and NWD into the Deep OC-SORT and OC-SORT methods with high-low-scores mechanism and verifies its effectiveness using the MOT20 dataset, which contains more low-quality targets with small sizes. The validation results are presented in Table 6. Compared to Deep OC-SORT, which utilized IoU originally, the employment of GIoU has achieved the improvement of 0.8, 3.56, 2.87, 2.05, and 316 in HOTA, MOTA, IDF1, AssA, and IDSW, but the employment of NWD has a more substantial enhancement 0.23, 0.31, 0.63, 0.45, and 83 in HOTA, MOTA, IDF1, AssA, and IDSW compared with GIoU. Similarly, for OC-SORT,

compared to the original IoU, the use of GIoU leads to the increment of metrics, but NWD shows the best performance.

Table 6: Ablation study on MOT20 with a single NWD

MOT20					
Tracker	HOTA↑	MOTA↑	IDF1↑	AssA↑	IDSW↓
Deep OC-SORT(IoU)	59.98	69.94	75.24	58.61	1182
Deep OC-SORT(GIoU)	60.78	73.50	78.11	60.66	866
Deep OC-SORT(NWD)	61.01	73.81	78.74	61.11	783
OC-SORT(IoU)	55.95	69.56	72.56	54.42	1297
OC-SORT(GIoU)	55.32	69.56	72.53	54.47	1320
OC-SORT(NWD)	56.03	69.54	72.56	54.55	1322

From the experimental results and analysis, it is easy to find that IoU is sensitive to positional deviations of the target. GIoU can the IoU by introducing a minimum bounding rectangle to represent the distance between two boxes, addressing the positional relationship when the bounding boxes do not overlap. However, GIoU remains reliant on IoU and defaults to IoU when the bounding boxes contain overlapping information. In contrast, NWD leverages Gaussian distribution to account for the overall distribution characteristics of the target area, and this approach can effectively measure the similarity even when there is limited overlap or mutual containment between bounding boxes, and it is less sensitive to scale. Consequently, NWD is more suitable for low-quality targets and demonstrates greater robustness in target association.

4.4 Parameter Sensitivity Study

(1) Threshold for high and low-quality targets selection.

In order to verify the influence of high and low score thresholds on data association, we only change the score thresholds τ_{low} and τ_{high} of the data association for the LQTTrack and conduct experiments on the MOT17 set to compare experiment results with different thresholds. We change τ_{low} from 0.1 to 0.3 and τ_{high} from 0.6 to 0.8, the results are shown in Table 7. As seen from the Table 7, when the low score threshold τ_{low} is 0.1, and the high score threshold τ_{high} is 0.6, LQTTrack achieves optimal performance. Therefore, the low and high score thresholds τ_{low} and τ_{high} in our algorithm are taken to be 0.1 and 0.6, respectively. Reasonable threshold selection has a positive impact on the data association, the suitable score thresholds significantly improve the association indicators, reducing trajectory fragmentation and identity switching.

Table 7: Comparative experiments of different thresholds in LQTTrack

MOT17					
Threshold	HOTA↑	MOTA↑	IDF1↑	AssA↑	IDSW↓
$\tau_{low} = 0.1$, $\tau_{high} = 0.6$	70.61	78.79	83.66	74.40	84

(Continued)

Table 7 (continued)

	MOT17				
Threshold	HOTA↑	MOTA↑	IDF1↑	AssA↑	IDSW↓
$\tau_{low} = 0.2$, $\tau_{high} = 0.7$	70.35	78.95	83.27	74.01	84
$\tau_{low} = 0.3$, $\tau_{high} = 0.8$	68.97	78.82	83.11	73.89	94

(2) Parameter for similarity computation.

The parameter adaptive weight a_w and dynamic appearance value λ_f are sensitive hyper-parameters and need to be carefully tuned in the task of multi-object tracking. Based on the experience of [10], we change a_w from 1.0 to 2.25 and λ_f from 0.75 to 1.0 and compare the HOTA, AssA, and IDSW of LQTTrack. The results are shown in Fig. 6. From the results, we can see that 1.25 and 0.95 are appropriate choices for a_w and λ_f. More suitable parameters can better weigh the appearance and motion features, improve matching accuracy, and enhance the effectiveness and performance of multi-target tracking. Therefore, we will also consider incorporating research on λ_f adaptive parameters in our future work, which will contribute to the advancement of our subsequent efforts.

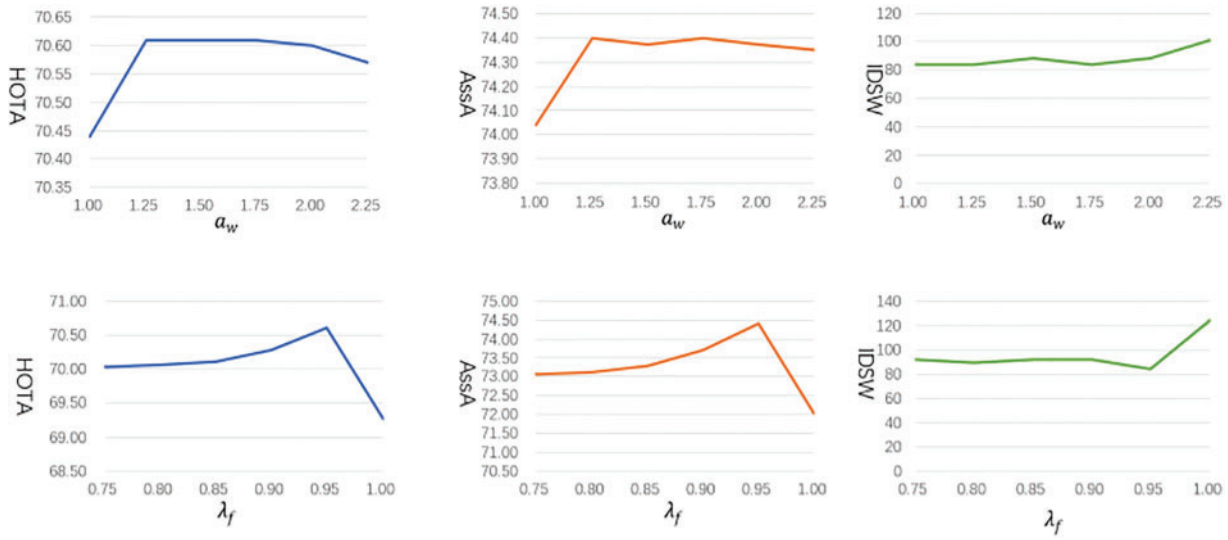

Figure 6: Study for the parameter of the adaptive weight a_w and dynamic appearance value λ_f on the MOT17 validation set

4.5 Runtime

The FPS is measured with NVIDIA GeForce RTX 3080Ti GPU. The FPS of the proposed is analyzed on the test set of MOT17 and MOT20, and the results are represented in Fig. 7. MSFF-FPN represents a multi-scale feature fusion module combined with the baseline, whereas LQA denotes a low-quality target association strategy without the MSFF-FPN, also combined with the baseline. For MOT17, from Fig. 7a, MSFF-FPN exhibits relatively lower operational efficiency, but LQA demonstrates a comparative advantage in operational efficiency, achieving increased FPS. Overall, LQTTrack enhances performance without imposing additional computational burdens. For MOT20,

from Fig. 7b, the effect of MSFF-FPN and LQA on FPS is similar to the MOT17 sets. However, LQTTrack exhibits relatively lower operational efficiency. Substantially, MSFF-FPN exhibits relatively lower operational efficiency, but LQA demonstrates a comparative advantage in operational efficiency, achieving increased FPS. This is primarily due to the increased computational complexity resulting from feature fusion while enhancing model performance and reducing processing speed. We will also focus on optimizing the efficiency of LQTTrack in our next phase of work.

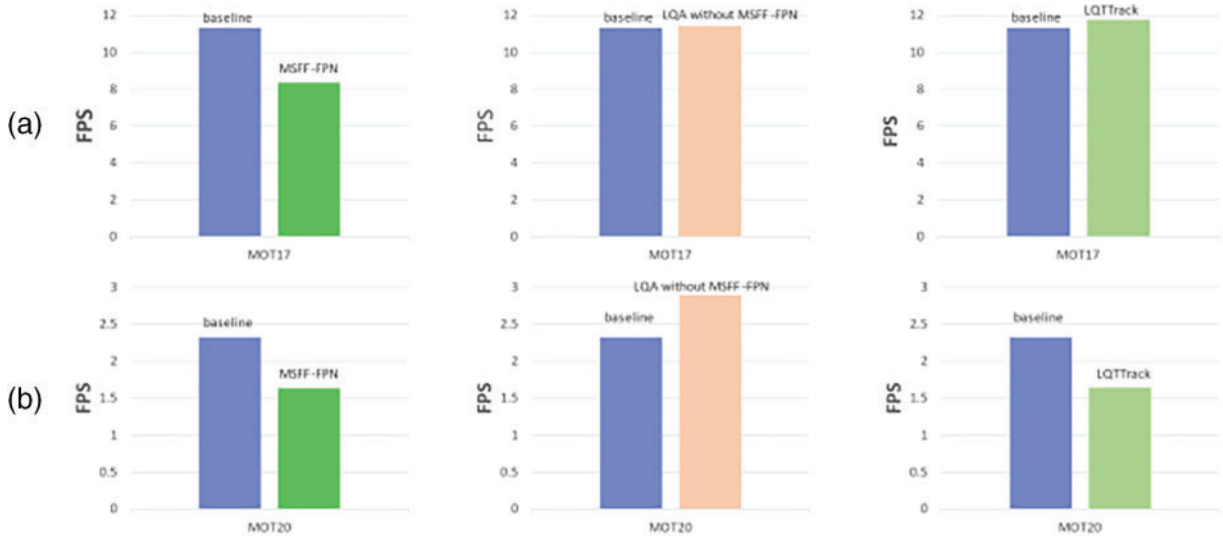

Figure 7: FPS of the method on the test sets of MOT17 and MOT20. (a) represents the FPS performance comparison of each module on the MOT17 dataset, (b) represents the FPS performance comparison of each module on the MOT20 dataset

5 Conclusion

In this work, we present an effective method LQTTrack for multi-object tracking with low-quality. LQTTrack is very effective in occlusion with the help of low-quality association schemes and NWD and feature fusion, enhancing the accuracy and robustness of multi-object tracking. We have verified the effectiveness of our method on MOT17, MOT20, and DanceTrack benchmarks. In the future, we will perform additional research on feature extraction and data association techniques to enhance the outcomes of multi-object tracking.

Acknowledgement: The authors would like to express our sincere gratitude and appreciation to each other for our combined efforts and contributions throughout the course of this research paper.

Funding Statement: This research was supported by the National Natural Science Foundation of China (No. 62202143) and Key Research and Promotion Projects of Henan Province (Nos. 232102240023, 232102210063, 222102210040).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Suya Li, Ying Cao; data collection: Xin Xie; analysis and interpretation of results: Hengyi Ren, Dongsheng Zhu; draft manuscript preparation: Suya Li, Ying Cao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data will be available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

[1] L. Jiao, D. Wang, Y. Bai, P. Chen, and F. Liu, "Deep learning in visual tracking: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5497–5516, 2023. doi: 10.1109/TNNLS.2021.3136907.

[2] S. Hassan, G. Mujtaba, A. Rajput, and N. Fatima, "Multi-object tracking: A systematic literature review," *Multimed. Tools Appl.*, vol. 83, no. 14, pp. 43439–43492, 2024. doi: 10.1007/s11042-023-17297-3.

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

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

[5] Y. Zhang *et al.*, "ByteTrack: Multi object tracking by associating every detection box," in *Comput. Vis.–European Conf. Comput. Vis. (ECCV) 2022*, Israel, 2022, pp. 1–21.

[6] Y. Du *et al.*, "StrongSORT: Make deepsort great again," *IEEE Trans. Multimed.*, vol. 25, pp. 8725–8737, 2023.

[7] J. Peng *et al.*, "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," in *Eur. Conf. Comput. Vis. 2020*, Glasgow, UK, 2020, pp. 145–161.

[8] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vision.*, vol. 129, no. 11, pp. 3069–3087, 2021. doi: 10.1007/s11263-021-01513-4.

[9] N. Aharon, R. Orfaig, and B. -Z. Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," 2022, *arXiv:2206.14651*.

[10] G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani, "Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification," in *2023 IEEE Int. Conf. Image Process. (ICIP)*, Kuala Lumpur, Malaysia, 2023, pp. 3025–3029.

[11] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 4696–4704.

[12] H. Sheng, J. Chen, Y. Zhang, W. Ke, Z. Xiong and J. Yu, "Iterative multiple hypothesis tracking with tracklet-level association," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3660–3672, 2019. doi: 10.1109/TCSVT.2018.2881123.

[13] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2117–2125.

[14] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized gaussian wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.

[15] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus and X. Alameda-Pineda, "TransCenter: Transformers with dense queries for multiple-object tracking," 2021, *arXiv:2103.15145*.

[16] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric SORT: Rethinking sort for robust multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, Vancouver, BC, Canada, 2023, pp. 9686–9696.

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

[18] P. Dendorfer *et al.*, "MOT20: A benchmark for multi object tracking in crowded scenes," 2020, *arXiv:2003.09003*.
[19] P. Sun *et al.*, "DanceTrack: Multi-object tracking in uniform appearance and diverse motion," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 20961–20970.
[20] Y. Dai, Z. Hu, S. Zhang, and L. Liu, "A survey of detection-based video multi-object tracking," *Displays*, vol. 75, no. 5, 2022, Art. no. 102317. doi: 10.1016/j.displa.2022.102317.
[21] Z. Pang, Z. Li, and N. Wang, "SimpleTrack: Understanding and rethinking 3D multi-object tracking," in *Comput. Vis.–ECCV 2022 Workshops, Tel Aviv*, Israel, 2022, pp. 680–696.
[22] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 1487–1495.
[23] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Republic of Korea, 2019, pp. 3701–3711.
[24] J. Seidenschwarz, G. Brasó, V. C. Serrano, I. Elezi, and L. Leal-Taixé, "Simple cues lead to a strong multi-object tracker," in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 13813–13823.
[25] D. Stadler and J. Beyerer, "Modelling ambiguous assignments for multi-person tracking in crowds," in *2022 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2022, pp. 133–142.
[26] K. Yi *et al.*, "UCMCTrack: Multi-object tracking with uniform camera motion compensation," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, 2024, pp. 6702–6710.
[27] H. -N. Hu, Y. -H. Yang, T. Fischer, T. Darrell, F. Yu and M. Sun, "Monocular quasi-dense 3D object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1992–2008, 2023. doi: 10.1109/TPAMI.2022.3168781.
[28] T. Fischer *et al.*, "QDTrack: Quasi-dense similarity learning for appearance-only multiple object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15380–15393, 2023. doi: 10.1109/TPAMI.2023.3301975.
[29] T. Keawboontan and M. Thammawichai, "Toward real-time UAV multi-target tracking using joint detection and tracking," *IEEE Access*, vol. 11, pp. 65238–65254, 2023. doi: 10.1109/ACCESS.2023.3283411.
[30] A. Psalta, V. Tsironis, and K. Karantzalos, "Transformer-based assignment decision network for multiple object tracking," *Comput. Vis. Image Understanding*, vol. 241, 2024, Art. no. 103957. doi: 10.1016/j.cviu.2024.103957.
[31] J. Hyun, M. Kang, D. Wee, and D. -Y. Yeung, "Detection recovery in online multi-object tracking with sparse graph tracker," in *2023 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2023, pp. 4839–4848.
[32] B. -T. Vo and B. -N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Trans. Signal Process*, vol. 61, no. 13, pp. 3460–3475, 2013. doi: 10.1109/TSP.2013.2259822.
[33] X. Xue, S. Huang, J. Xie, J. Ma, and N. Li, "Resolvable cluster target tracking based on the DBSCAN clustering algorithm and labeled RFS," *IEEE Access*, vol. 9, pp. 43364–43377, 2021. doi: 10.1109/ACCESS.2021.3066629.
[34] L. Van Ma, T. T. D. Nguyen, C. Shim, D. Y. Kim, N. Ha and M. Jeon, "Visual multi-object tracking with re-identification and occlusion handling using labeled random finite sets," *Pattern Recognit.*, vol. 156, no. 13, 2024, Art. no. 110785. doi: 10.1016/j.patcog.2024.110785.
[35] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 658–666.
[36] J. Kong, E. Mo, M. Jiang, and T. Liu, "MOTFR: Multiple object tracking based on feature recoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7746–7757, 2022. doi: 10.1109/TCSVT.2022.3182709.

[37] W. Feng, L. Lan, Y. Luo, Y. Yu, X. Zhang and Z. Luo, "Near-online multi-pedestrian tracking via combining multiple consistent appearance cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1540–1554, 2021. doi: 10.1109/TCSVT.2020.3005662.

[38] H. Zhang *et al.*, "ResNeSt: Split-attention networks," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New Orleans, LA, USA, 2022, pp. 2735–2745.

[39] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu and G. -S. Xia, "Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, no. 9, pp. 79–93, 2022. doi: 10.1016/j.isprsjprs.2022.06.002.

[40] E. Yu, Z. Li, S. Han, and H. Wang, "RelationTrack: Relation-aware multiple object tracking with decoupled representation," in *IEEE Trans. Multimed.*, 2023, vol. 25, pp. 2686–2697.

[41] Q. Wang, Y. Zheng, P. Pan, and Y. Xu, "Multiple object tracking with correlation learning," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 3875–3885.

[42] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "TransMOT: Spatial-temporal graph transformer for multiple object tracking," in *2023 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2023, pp. 4859–4869.

[43] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP J. Image Video Process*, vol. 2008, pp. 1–10, 2008. doi: 10.1155/2008/246309.

[44] J. Luiten *et al.*, "HOTA: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vision.*, vol. 129, no. 2, pp. 548–578, 2021. doi: 10.1007/s11263-020-01375-2.

[45] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Comput. Vis.–ECCV 2016 Workshops*, Amsterdam, The Netherlands, 2016, pp. 17–35.

[46] L. He, X. Liao, W. Liu, X. Liu, P. Cheng and T. Mei, "FastReID: A pytorch toolbox for general instance re-identification," 2020, *arXiv:2006.02631*.

[47] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Comput. Vis.–ECCV 2020*, Glasgow, UK, 2020, pp. 474–490.