Image and Vision Computing 162 (2025) 105695

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Synergistic-aware cascaded association and trajectory refinement for multi-object tracking

Hui Li, Su Qin, Saiyu Li, Ying Gao *, Yanli Wu

School of Data Science, Qingdao University of Science and Technology, Qingdao, 266061, China

ARTICLE INFO

Keywords:
Computer vision
Multi-object tracking
Multi-frame collaboration
Confidence

ABSTRACT

Multi-object tracking (MOT) is a pivotal research area in computer vision. Effectively tracking objects in scenarios with frequent occlusions and crowded scenes has become a key challenge in MOT tasks. Existing tracking-by-detection (TbD) methods often rely on simple two-frame association techniques. However, in situations involving scale transformation or requiring long-term association, frequent occlusion between objects can lead to ID switches, especially in scenes with dense or highly intersecting objects. Therefore, we propose a synergistic-aware cascaded association and trajectory refinement method (SCTrack) for multi-object tracking. In the data association stage, we propose a synergistic-aware cascaded association method to construct a multi-perception affinity matrix for object association, and introduce the multi-frame collaborative distance calculation to enhance the robustness. To address the problem of trajectory fragmentation, we propose a dynamic confidence-driven trajectory refinement post-processing method. This method integrates confidence and feature information to calculate trajectory association, repair fragmented trajectories, and improve the overall robustness of the tracking algorithm. Extensive experiments on the MOT17, MOT20, and DanceTrack datasets validate SCTrack's competitive performance.

1. Introduction

In the contemporary realm of computer vision, multi-object tracking (MOT) [1–7] is a task of paramount importance that continues to garner widespread attention. With the rapid development of intelligent technology, the significant potential of MOT is gradually becoming evident across various domains. Its extensive applications in areas such as video surveillance [8], intelligent transportation [9], human–computer interaction [10], and autonomous driving [11] provide robust support for constructing intelligent systems, enhancing the quality of life, and addressing real-world challenges.

As a result, the predominant paradigm remains track-by-detection (TbD). This paradigm first employs object detectors to identify objects in the scene. Then, using motion, position, appearance, or their combinations, it associates the detection results across frames to form trajectories corresponding to specific identities. The advantage of the TbD paradigm is its relatively efficient computation, which improves the efficiency of the tracking algorithm by improving the performance of the detector. Existing MOT methods still share some common shortcomings when dealing with complex scenarios:

(1) Frequent ID Switch. Frequent identity switching is a challenging issue in MOT, especially in scenarios with dense or highly intersecting objects. Newly proposed datasets, such as the DanceTrack [12] datasets, feature multiple objects with high similarity, which can lead to ID switches. In Fig. 1(a), we observe that in scenarios with similar appearance characteristics, an ID switch occurs when Object 4 reappears after overlapping with Object 2, resulting in association errors. In complex and highly dynamic scenarios, object features may change drastically, leading to matching failures. Fig. 1(b) shows an ID switch between Object 2 and Object 18. Traditional feature extraction methods struggle to adapt to the dynamic changes of objects, failing to accurately capture key information, which weakens the discriminative power of the affinity matrix and impacts the robustness of the entire tracking algorithm.

(2) Trajectory Fragmentation. Trajectory post-processing is a weak link in MOT. Some methods only use simple appearance features for track refinement, while others guarantee track continuity through offline interpolation. In the process of object movement, frequent occlusion and instantaneous motion blur will lead to tracking failure. Fig. 1(c) shows that after frequent occlusions, Object 8 reappears and is initialized as a new trajectory. In fact, Object 52 and Object 8 are part of the same trajectory, but the tracker fails to consistently track the object, leading to trajectory fragmentation. In complex scenes, it is particularly challenging to accurately reconstruct the complete trajectory

* Corresponding author.
E-mail address: gaoying@qust.edu.cn (Y. Gao).

https://doi.org/10.1016/j.imavis.2025.105695
Received 19 August 2024; Received in revised form 30 June 2025; Accepted 3 August 2025
Available online 23 August 2025

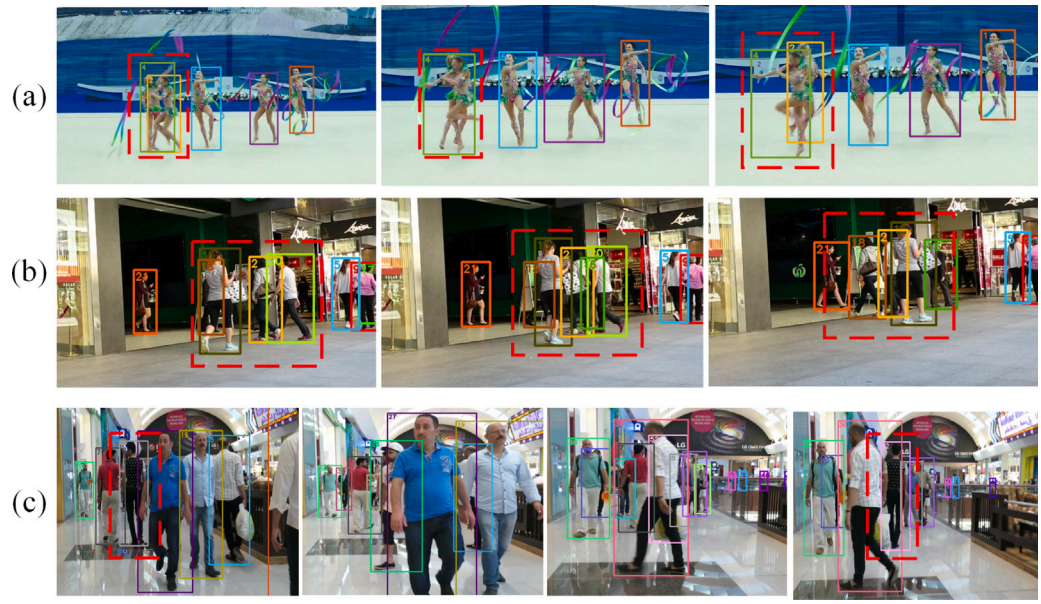

Fig. 1. Challenges and Difficulties in MOT. (a) The red dashed line shows that after Object 2 obscures Object 4, due to their similar appearances, an ID switch occurs when Object 4 reappears. (b) In the red dashed line, an association error between Object 2 and Object 18 is shown under crowded conditions and frequent occlusion. (c) The red dashed line indicates that after a long period of being obscured, Object 8 reappears and its trajectory is initialized as a new trajectory.

of frequently occluded objects, affecting the overall performance of the tracking algorithm.

To address these challenges, we propose a synergistic-aware cascaded association and trajectory refinement method (SCTrack) for multi-object tracking. This approach enhances synergistic-aware cascaded association and trajectory refinement to significantly improve the performance of MOT, better adapting to complex and dynamic real-world scenarios.

To summarize, our contributions are listed as below:

- We propose Synergy-aware Cascaded Association module in the data association stage. This module includes a Multi-Directional Feature Fusion Attention mechanism to construct multi-perception affinity matrix and utilizes Multi-frame Collaborative Distance Calculation to enhance matching precision.
- To address the issue of trajectory fragmentation, we propose a Dynamic Confidence-Driven Trajectory Refinement module. This module dynamically calculates trajectory associations by integrating confidence scores with appearance features, thereby repairing trajectories and ensuring their continuity.
- Our method is successfully applied to the MOT17, MOT20, and DanceTrack datasets, achieving excellent performance.

2. Related work

The TbD paradigm is one of the most popular methods for MOT, decomposing the task into two subtasks: first, using object detection methods [13–21] to acquire bounding boxes, and second, associating objects across different frames through various methods. Current MOT methods predominantly emphasize the data association stage.

2.1. Data association

Single-cue Data Association. Early MOT methods primarily rely on either motion or appearance features to associate objects across frames.

Motion-based approaches use object movement prediction to establish associations. Xiao et al. [22] introduced a learnable motion predictor to capture complex motion patterns and handle long-term occlusions, incorporating both an interaction module for modeling object interactions and a rediscovery module for re-identifying lost objects, achieving state-of-the-art results. Zhang et al. [23] proposed a two-stage association method that discards appearance features, dividing detections into high-score and low-score categories. Qin et al. [3] designed a rediscovery mechanism that re-links unmatched trajectories based on motion continuity. Cao et al. [2] tackled long-term occlusion by generating virtual trajectories derived from detector observations to correct accumulated filtering errors. However, motion-only methods often show limited robustness in complex or highly dynamic scenes, particularly under long-term occlusion or abrupt motion changes.

Appearance-based approaches, on the other hand, leverage deep visual embeddings to enhance association accuracy, especially under occlusion or irregular motion. Huang et al. [24] proposed a non-motion-based MOT framework that uses only high-quality detection and ReID features, achieving strong performance. Aharon et al. [4] and Wang et al. [25] further explored visual similarity metrics using feature extraction and self-attention mechanisms. Despite their effectiveness, appearance-only methods can incur high computational costs and are prone to confusion in datasets containing visually similar objects, such as the DanceTrack dataset.

Multi-cue Fusion Methods. To improve the robustness and adaptability of MOT algorithms, recent studies have increasingly explored combining motion, appearance, and auxiliary cues for more reliable data association. For example, Liu et al. [26] proposed a pseudo-depth estimation method that derives relative depth information from 2D images to reduce occlusion-related errors. This method demonstrates strong potential in both occlusion handling and practical deployment, although challenges remain when tracking fast-moving objects. Maggiolino et al. [27] presented a substantial extension of prior work [2] by incorporating visual appearance features, camera motion compensation, and dynamically adjusted embeddings, along with a weighting strategy to refine the cost matrix. This integration significantly improved tracking robustness and accuracy. Other methods further enhance appearance modeling: Wang et al. [25] employed a siamese network with self-attention to extract more discriminative features; and Jin et al. [28] introduced a feature-decoupled framework to suppress noisy or redundant information, reducing computational overhead. Seidenschwarz et al. [29] proposed a TbD tracker that distinguishes active from inactive trajectories and adjusts distance metrics accordingly. Unlike traditional approaches that rely only on motion and appearance, our method also incorporates position cues and introduces a

multi-directional fusion attention to enhance feature representation. By leveraging multi-frame context and improving spatial-semantic discriminability, our approach enables more robust and consistent associations in complex tracking scenarios.

2.2. Trajectory restoration

Trajectory refinement is a crucial post-processing module in multi-object tracking systems, designed to handle trajectory interruptions caused by occlusion, frame drops, or object detection failures. To address this issue, the refinement process jointly models temporal consistency, spatial continuity, and appearance similarity between fragmented tracklets in order to reconnect those that belong to the same object. This enables the recovery of the object's continuous motion path throughout the entire video sequence. Du et al. [30] proposed a novel trajectory refinement module that completely discards traditional reliance on appearance features. Instead, it focuses on spatial relationships between tracklets and performs Gaussian smoothing interpolation based on the coordinates of the top-left corners of bounding boxes. While this method achieves promising results, it depends heavily on batch processing and lacks the ability to support real-time or online applications effectively.

Feature-matching-based trajectory reconnection strategies aim to re-link fragmented trajectories by evaluating the similarity between their features to determine whether they belong to the same object. Wang et al. [31] computed the appearance features of objects and simultaneously evaluated the trajectory similarity by calculating the latest 10 person ReID (Re-Identification) pairs, thus achieving trajectory repair through merging trajectories. However, this method has high computational complexity, and its accuracy in trajectory recovery decreases significantly under long-term occlusion. Wang et al. [32] proposed a trajectory splitting strategy, which splits unreasonable trajectories by using appearance or motion information. Although this provides a new approach for multi-object tracking, it can lead to false detections when the object's appearance changes significantly. Cao et al. [2] employed the estimation of object centroids for restoration. Du et al. [30] introduced an offline trajectory restoration module that abandons appearance information, associates short trajectories into complete ones, and utilizes Gaussian smoothing interpolation based on the top-left corner coordinates. However, this method is run offline, exhibits high computational complexity, and is not suitable for real-time applications. These methods indicate that researchers are continuously refining and enhancing the robustness and accuracy of MOT. However, further research and improvements in computational complexity and real-time performance are still needed. Unlike traditional refinement methods that rely on rigid interpolation or fixed appearance similarity, our approach adaptively recovers fragmented trajectories by jointly considering confidence trends and discriminative visual cues. This enables more reliable trajectory continuity in the presence of long-term occlusion and dense target interactions.

3. Method

Our overall roadmap is illustrated in Fig. 2, primarily comprising two modules: the Synergistic-aware Cascaded Association and the Dynamic Confidence-Driven Trajectory Refinement module. The Synergistic-aware Cascaded Association module is further divided into two sections for detailed description: Section 3.1 describes the construction of the Multi-Perception Affinity Matrix (MPAM), and Section 3.2 discusses the Multi-frame Collaborative Distance Calculation (MCDC). Section 3.3 elaborates on the Dynamic Confidence-Driven Trajectory Refinement (DCD) module.

3.1. Multi-perception affinity matrix

The inclusion of rich embedded feature information significantly enhances the robustness of tracking algorithms in MOT. However, capturing discriminative features in dynamically changing environments poses challenges for feature extraction networks, which may struggle to differentiate object features effectively. This limitation often results in insufficient feature extraction, hampering the tracking of objects in complex scenarios like frequent occlusions and crowded environments. To address this, we have designed a feature extraction network that comprehensively integrates object position, motion, and appearance features. Unlike traditional methods that consider only appearance and motion cues, our network further incorporates explicit positional features such as object velocity and directional angle. A Multi-Directional Fusion Attention mechanism is employed to adaptively capture interactions across spatial and channel dimensions. The network architecture is illustrated in Fig. 3.

After obtaining the F_0 feature map, P_1 and F_1 are generated separately through convolutional neural networks. Subsequently, the processing flow enters two branches. In one branch, the processing builds upon F_1 and introduces dilated convolutions to extract multi-level information from the feature map. Dilated convolutions have the advantage of incorporating features from multiple mappings of different regions, comprehensively capturing information in the image. To maintain feature map invariance, deformable convolutional layers are introduced in each path, enhancing the robustness and generalization capabilities of the module. In the other branch, convolution and element-wise addition operations are performed with F_3. These two branches respectively provide coarse-grained and fine-grained information, enabling the network to comprehensively understand and express features of the input data. By combining information from these two branches, the network achieves multi-scale feature learning, making the model more adaptable to complex data structures and variations. This network design excels in capturing both local and global information when handling image tasks, enhancing overall performance and effectiveness. The formula for computing F_2 is as follows formula (1):

$$F_2 = concat(dilated(x), deform(x)) \tag{1}$$

where the input for F_1 is denoted as x, $dilated(x)$ represents dilated convolution, and $deform(x)$ stands for deformable convolution. We took inspiration from the paper [33], and sequentially set the dilation rates for the dilated convolutions in each path as 3, 6, 12, and 18. The results are then concatenated to form F_2, which captures rich contextual information. To integrate with P_1, we generated F_3 through a 1×1 convolution and performed the addition operation. This yields more enriched semantic information and enhances the performance of the tracking algorithm.

However, not all information obtained, even if F_4 contains rich feature details, may be useful for the subsequent data association process due to the presence of noise. To address this issue, we propose a technique named Multi-Directional Feature Fusion Attention (MDF-Fusion Attention) to enhance the features. Our method utilizes attention mechanisms in different directions, namely vertical, horizontal, and channel, to process context information separately. Subsequently, we fuse the extracted context information with information from different directions to effectively capture valuable information, resulting in the final feature F_5.

First, after obtaining the feature map F_4, we divide the input x into g groups based on the number of channels c. In other words, we represent x as $x = [x_0, \ldots, x_w, x_{g-1}]$, where $x_w \in x$, and the size of x_w is $(b \times g, c/g, height, width)$. We sequentially input these grouped features into subnets and extract attention weights for the grouped features through three parallel pathways. To introduce coordinate attention, we use two branches in the parallel paths, as shown in the formula (2):

$$\begin{aligned} w(x_w) &= X\_adjust\_avgpool(x_w) \\ h(x_w) &= Y\_adjust\_avgpool(x_w) \end{aligned} \tag{2}$$

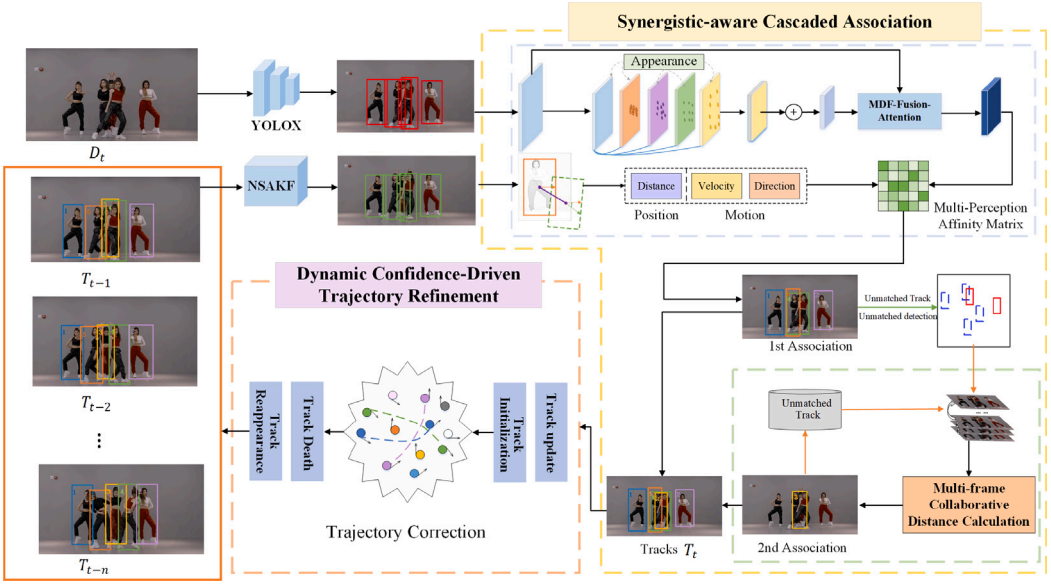

Fig. 2. Overview of SCTrack. SCTrack includes a cascaded association module and a trajectory refinement module. It extracts multi-perception features from frame D_t, applies MDF-Fusion Attention, and uses multi-frame data to refine associations based on object confidence.

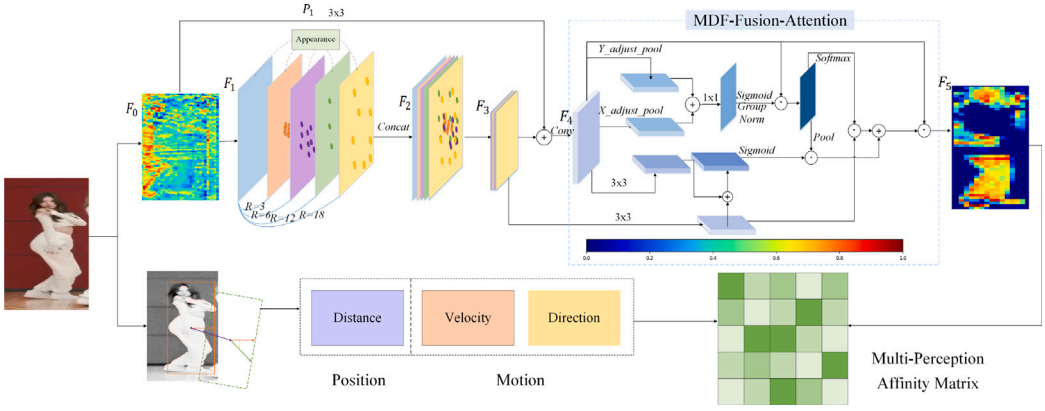

Fig. 3. Multi-Perception Affinity Matrix. F_0-F_5 and P_1 represent feature map. We extract the position, motion, and appearance information of the video frame, and construct a multi-perception affinity matrix. Additionally, the MDF-Fusion-Attention is proposed to enrich appearance information.

Among them, $h(x_w)$ and $w(x_w)$ represent adaptive average pooling operations in the vertical and horizontal directions respectively. We then fuse them across channels using a 1×1 convolutional to obtain hw. Next, we separate hw back into $h(x_w)$ and $w(x_w)$. We apply the *Sigmoid* function to $h(x_w)$ and $w(x_w)$, and then multiply the results element-wise with the original x_w. Finally, we normalize the results, as shown in the formula (3):

$$x_{w1} = GroupNorm(Sigmoid(h(x_w)) \odot x_w \odot Sigmoid(w(x_w))) \tag{3}$$

In the third path on the 3×3 branch, we introduce channel attention, denoted as x_{c1}. To leverage richer contextual information, we also introduce the F_3 feature map to channel attention, denoted as x_{c2}. We then concatenate these features with the features obtained from the third path. Next, we process x_{c1} and x_{c2} through global average pooling and the *Sigmoid* function to obtain attention weights, as shown in formula (4):

$$\begin{aligned} x_{w1} &= Softmax(x_{w1}) \odot Conv(x_{c2}) \\ x_c &= (Sigmoid(x_{c1} + x_{c2})) \odot Avgpool(x_{w1}) \end{aligned} \tag{4}$$

By performing the addition operation on the obtained x_c and x_{w1}, we perform an element-wise addition operation to obtain the attention matrix weights, as shown in the formula (5):

$$weights = x_c + x_{w1} \tag{5}$$

Finally, the *weights* are multiplied element-wise with the original input to generate the final output, as shown in formula (6):

$$x_w = x_w \times weights \tag{6}$$

After constructing the appearance feature affinity matrix, we introduce position and motion information by calculating the object velocity, orientation, and Euclidean distance to construct the velocity affinity matrix, orientation affinity matrix, and Euclidean distance affinity matrix. The sum of these three matrices is collectively referred to as the dynamic spatial affinity matrix. Assuming the velocity of the bounding box u and the predicted box v are V_u and V_v, the orientation angles are Θ_u and Θ_u, and the center points are (X_u, Y_v) and (X_u, Y_v), the specific calculations are shown in the formula (7):

$$\begin{aligned} Velocity(u,v) &= exp^{-|V_u - V_v|^2} \\ Direction(u,v) &= cos(\Theta_u - \Theta_v) \\ Euclidean(u,v) &= exp^{-\sqrt{(X_u - X_v)^2 + (Y_u - Y_v)^2}} \end{aligned} \tag{7}$$

where the exponential function in this context serves to map the differences in velocity and Euclidean distance to a similarity measure between 0 and 1. The dynamic spatial affinity matrix is weighted and summed with the appearance affinity matrix to obtain the final Multi-Perception Affinity Matrix (MPAM).

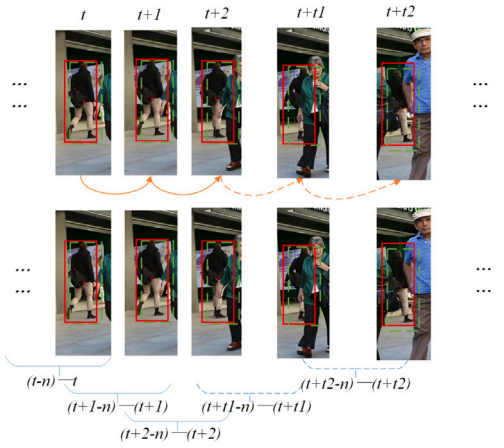

Fig. 4. Multi-frame Collaboration Strategy. The red box indicates the bounding box, while the green dashed box represents the predicted box. Traditional tracking methods only use two frames for simple IOU association but have a poor effect on long-term association. In our method, we introduce the concept of "Multi-Frame Collaboration", which involves considering the historical first n frames for data association.

3.2. Multi-frame collaborative distance calculation

After obtaining rich embedded information, data association becomes the pivotal step in object tracking, directly influencing the accuracy and precision of the tracking algorithm. Over-reliance on appearance information or motion information in data association can render it vulnerable to noise and errors. The former is particularly prone to the influence of image noise, while the latter may be affected by errors in motion estimation or the limitations of imprecise motion models. These issues can disrupt or misjudge object motion characteristics during association, thereby compromising accuracy and stability. To tackle these challenges, we aim to improve the accuracy and precision of the association stage by leveraging the multi-level information obtained earlier. Inspired by ByteTrack [23], which employs a two-stage association method, we go one step further by introducing a Multi-frame Collaborative Distance Calculation (MCDC) module. Unlike conventional two-frame matching, our MCDC module constructs the association matrix by computing a weighted sum between the historical frames of trajectories and the current detection frame. This history-aware strategy fully exploits long-term temporal cues, enhancing association robustness in complex and dynamic scenes. Fig. 4 illustrates the multi-frame collaboration strategy. By retaining information for up to 30 frames in memory and utilizing multi-frame historical data for association, our approach enhances algorithmic robustness.

First Association. To comprehensively consider the appearance, motion, and position information of objects, we construct a multi-perception affinity matrix. To further enhance the accuracy and stability of the object state during the association process, we adopt the NSA kalman filter proposed by GIAOTracker [34] for object prediction and updates. This filter incorporates confidence factors into the motion prediction and update procedures, thereby improving the accuracy and stability of object state estimation and enhancing the robustness of the association process.

Second Association. Fig. 5 illustrates our proposed MCDC module for the association process. Each trajectory consists of detections from each frame. The memory module retains up to 50 frames of lost trajectories, denoted as T_{lost}. Assuming the feature set of detections is $D_d = \{D_1, D_2, \ldots, D_j\}$, $1, 2, \ldots, j \in d$, and the feature set of trajectories is $T_q = \{T_1, T_2, \ldots, T_i\}$, $1, 2, \ldots, i \in q$. The feature of trajectory i is denoted as T_i, consisting of the appearance feature vectors from n frames, specifically $T_i = \{D_i^1, D_i^2, \ldots, D_i^n\}$. Then, we calculate the similarity distance between trajectory T_i in the tth frame and a new detection D_j. The appearance information for each detection in D_i^n is denoted as $D_i^n = \{f_i^1, f_i^2, \ldots, f_i^N\}$. The cosine similarity $\xi(D_i^n, D_j)$ is as shown in formula (8):

$$\xi(D_i^n, D_j) = \frac{\sum_{l=1}^{N} f_i^l \cdot f_j^l}{\sum_{l=1}^{N} (f_i^l)^2 \cdot \sum_{l=1}^{N} (f_j^l)^2} \tag{8}$$

where N represents the dimensionality of the features. Then, we calculate the average of these distances, considering it as the final distance measure $\Gamma(T_i, D_j)$, as specified in the formula (9):

$$\Gamma(T_i, D_j) = \frac{1}{n} \sum_{k=1}^{n} \xi(D_i^k, D_j^t) \tag{9}$$

Based on this method, we implement multi-frame collaboration in the second association to improve the precision and relevance of matching. Finally, trajectories that do not find matches in both association stages are retained. If a trajectory remains unassociated for 30 consecutive frames, it is released. Newly detected items that persist for more than five frames are initialized as new trajectories, while associated trajectories proceed to the matching process in the subsequent frame.

3.3. Dynamic confidence-driven trajectory refinement

Trajectory post-processing plays a crucial role in optimizing tracking results and maintaining temporal coherence in MOT. To this end, we propose a Dynamic Confidence-Driven Trajectory Refinement (DCD) module, which dynamically adjusts the association strength between trajectories based on detection confidence. Unlike conventional approaches that rely on fixed similarity thresholds, our method leverages confidence scores to guide both trajectory association and management decisions—a strategy rarely explored in existing literature. This enables more adaptive and reliable trajectory refinement, particularly in complex or low-confidence scenarios. We outline the pseudocode for this module in Algorithm 1. The inputs include a video sequence S, a set of detections D_d, and a trajectory association threshold δ. Lines 2–4 describe the process of data association twice, updating the associated trajectory sets T_{q1} and T_{q2} along with high-confidence unmatched detections D_{undet}. Lines 5–11 handle trajectory management, where the formula (10) for calculating the trajectory features in the frame k is as follows:

$$T_i.feats = \frac{1}{n} \sum_{k=1}^{n} (c_i^k \cdot D_i^k.feat) \tag{10}$$

where c_i^k represents the detection confidence of trajectory i in frame k, and $D_i^k.feat$ denotes the appearance features of trajectory i in the same frame, where $i \in \{1, 2, \ldots, q\}$. The feature of trajectory i in frame k, The feature of trajectory $T_i^k.feats$, is calculated by weighting the confidence and appearance features together and performing a summation. This results in the final feature of trajectory i. We calculate the similarity between trajectories T_1 and T_2 using cosine similarity, as shown in the formula (11):

$$TrackCore(T_1, T_2) = \xi(T_1.feats, T_2.feats) \tag{11}$$

Finally, we make a repair judgment based on the given threshold δ on the calculated trajectory similarity. If the similarity exceeds the threshold, we consider the two trajectories to be correlated and proceed with trajectory restoration. Trajectory restoration employs the Gaussian interpolation method mentioned in [2,30], which smoothens trajectories to enhance coherence and optimize overall tracking performance. This comprehensive trajectory post-processing method effectively enhances the performance of the MOT algorithm, making it suitable for various practical application scenarios. Line 11 removes long-term unmatched trajectories, denoted as T_{lost}, and outputs the final trajectory set T_q.

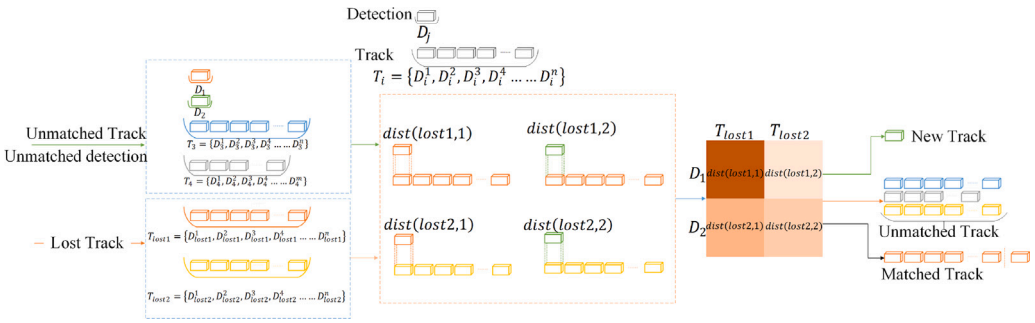

Fig. 5. Multi-frame Collaborative Distance Calculation. We calculate the similarity between the appearance feature vectors of T_i and a new detection D_j in the tth frame as the final distance measure.

Algorithm 1: Dynamic Confidence-Driven Trajectory Refinement

input : Input:Video sequence S ;Detections D;
Track association threshold δ;
output: Track T_q

Initialization: $T_q \leftarrow \emptyset$
for *frame* S_{frame} *in* S **do**
  Association1:Trajectory set $T_{q1} \leftarrow Matched\ Track$;
  Association2:Trajectory set $T_{q2} \leftarrow Matched\ Track$;
    $T_{lost} \leftarrow Unmatched\ \ Track$;
    $D_{undet} \leftarrow Unmatched\ \ Dectections$;
  $T_q = T_{q1} \cup T_{q2} \cup D_{undet}$; // Track update
  // Track management
  for $T_1.feats$ *in* T_q **do**
    for $T_2.feats$ *in* T_q **do**
      // Compute trajectory similarity
      if $T_1.feats! = T_2.feats$ *and*
       $TrackCore(T_1.feats, T_2.feats) > \delta$ **then**
      // Gaussian interpolation
      $GTR(T_1, T_2)$
      end
    end
  end
  // Delete unmatched tracks
  $Delete(T_{lost})$
end
Return T_q

4. Experiments

4.1. Implementation details

In our experiments, we utilize the FastReID [35] framework as the training network, conducting training over 60 epochs in accordance with the training protocol described in [4]. Throughout the entire training process, we maintain consistent parameters to ensure uniformity and reproducibility of experimental conditions. For object detection, we utilize a private detection method based on the YOLOX [13] detector. We obtain detection results following the procedures outlined in [23]. The experimental hardware configuration is detailed in Table 1.

4.2. Datasets and metrics

Datasets. To evaluate our method, we utilize three public datasets: MOT17, MOT20, and DanceTrack. These datasets cover a wide range of challenging scenarios including frequent occlusions, lighting variations, and similar appearances, providing a comprehensive evaluation of our method.

Table 1
Experimental hardware configuration.

Configuration	Parameters
OS	Ubuntu 20.04.3 LTS
CPU	Intel(R) Core(TM) i7-11700F @ 2.50 GHz
GPU	NVIDIA Corporation GP102 [GeForce GTX 1080 Ti] (rev a1)
Python	3.7.12
Pytorch	1.8.0
Torchvision	0.9.0
CUDA	11.1.1
CUDNN	8.0.5

MOT17. The public datasets are widely used in the field of MOT, and they typically consist of multiple subsets, each providing large-scale video sequences. These sequences include realistic scenes such as high-density crowds and complex intersections, among other challenging scenarios.

MOT20. As a recent extension to the MOT series, building upon MOT17 by introducing more challenging scenarios and additional annotation information. This expansion facilitates the evaluation of MOT performance in diverse real-world applications.

DanceTrack. Some datasets in the MOT field focus on human dance behavior analysis and provide rich pose and motion information. DanceTrack presents a significant challenge for MOT algorithms due to the highly similar appearances of the dancers.

Metrics. When evaluating the performance of MOT algorithms, we employ a comprehensive set of evaluation metrics that assess tracking performance from various perspectives. These metrics include various parameters from literature [36], metrics suites such as MOT (Multi-Object Tracking Accuracy), FP (False Positives), FN (False Negatives), and IDS (IDentity Switch), along with IDF1 (IDentity F1 Score), HOTA (Higher Order Tracking Accuracy), MT (Mostly Tracked objects) and ML (Mostly lost objects). By using these metrics, we can gain a comprehensive evaluation of the capacity of the MOT algorithm to track multiple objects accurately and to manage challenges such as identity switches in various tracking scenarios.

4.3. State-of-the-art comparison

To further demonstrate the effectiveness of our module, we conduct experiments on the MOT17, MOT20, and DanceTrack test datasets, achieving optimal tracking results across multiple evaluation metrics. The Fig. 6 illustrates a comparison of MOT17 with advanced algorithms, evaluating based on IDF1 and MOTA metrics. The chart clearly shows that our approach has achieved state-of-the-art performance.

On the MOT17 test datasets, as shown in Table 2, our method achieves optimal performance in MOTA, HOTA, and IDF1. Compared to the baseline model, our approach improves HOTA by 0.7 and IDF1 by 0.6. In addition to these core indicators, our method also achieves significant results on other key performance indicators. Specifically, FP and FN rank second and third, respectively. It is noteworthy that

Table 2
MOT17 test datasets. **Black bold** ranks first. Compared with the state-of-the-art algorithms, SCTrack ranks first in MOTA, HOTA, and IDF1.

Method	Ref.	MOTA↑	HOTA↑	IDF1↑	FN↓	FP↓	MT↑	ML↓	IDS↓
FairMOT [37]	IJCV 2021	73.7	59.3	72.3	117,477	27,507	43.2%	17.3%	3303
CorrTracker [38]	CVPR2021	76.5	60.7	73.6	99,510	29,808	47.6%	12.7%	3369
MOTR [39]	ECCV2022	73.4	57.8	68.6	–	–	50.3%	13.1%	2439
P3AFormer [40]	ECCV2022	78.1	–	76.4	86,510	25,413	**70.5%**	**7.4%**	**1332**
MeMOT [41]	CVPR2022	72.5	56.9	69.0	115,248	37,221	43.8%	18.0%	2724
QDTrack [42]	TPAM2023	78.7	**63.5**	77.5	**83,154**	34,923	54.0%	12.6%	1935
OC-SORT [2]	CVPR2023	78.0	63.2	77.5	107,055	**15,129**	–	–	1950
SGT [43]	WACV2023	76.4	–	72.8	102,885	25,974	48.0%	11.7%	4101
CLNet [44]	ICASSP2024	72.0	57.8	72.2	137,622	17,790	37.6%	19.9%	2823
Baseline [29]	CVPR2023	78.7	62.8	77.1	–	–	–	–	2325
SCTrack	–	**79.4**	**63.5**	**77.7**	91,927	22,032	50.1%	16.3%	2022

Table 3
MOT20 test datasets. **Black bold** ranks first. Compared with the state-of-the-art algorithms, SCTrack achieves rankings in HOTA, IDF1, IDS, and FP.

Method	Ref.	MOTA↑	HOTA↑	IDF1↑	FN↓	FP↓	MT↑	ML↓	IDS↓
SORMOT [45]	CVPR2021	68.6	57.4	71.4	101,154	57,064	64.9%	9.7%	4209
CrowedTrack [38]	AVSS2021	70.7	55.0	68.2	126,533	21,928	54.9%	12.1%	3198
ByteTrack [23]	ECCV2022	**77.8**	61.3	75.2	**87,594**	26,249	**69.2%**	9.7%	1223
CSTrack [46]	TIP2022	66.6	–	68.6	144,358	25,404	50.4%	15.5%	3196
MAATrack [47]	WACV2022	73.9	57.3	71.2	108,744	24,942	59.7%	12.3%	1331
MeMOT [41]	CVPR2022	63.7	54.1	66.1	137,983	47,882	57.5%	14.3%	1938
FDTrack [28]	IEEE2023	75.0	59.9	75.7	102,896	24,011	62.8%	9.7%	2226
SGT [43]	WACV2023	72.8	–	70.6	112,963	25,161	64.3%	12.7%	2474
QDTrack [42]	TPAM2023	74.7	60.0	73.8	106,313	23,352	64.2%	13.0%	1043
CLNet [44]	ICASSP2024	65.6	54.9	68.4	96,990	75,731	66.7%	**8.4%**	5472
Baseline [29]	CVPR2023	73.7	61.2	75.2	–	–	–	–	1264
SCTrack	–	75.6	**61.4**	**76.1**	107,894	**17,779**	64.3%	12.8%	**837**

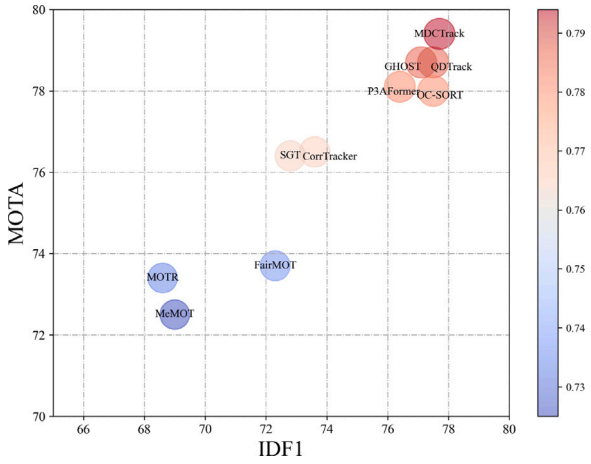

Fig. 6. Comparison of different trackers in MOTA-IDF1 test datasets for MOT17. The horizontal axis is IDF1, the vertical axis is MOTA. SCTrack achieves the most advanced results in MOTA and IDF1, as detailed in Table 2.

our MT and ML indicators are lower than those of QDTrack. The QDTrack algorithm achieves this by densely matching hundreds of image regions for contrastive learning, effectively capturing diverse object appearances. It utilizes a bi-directional $Softmax$ similarity metric to enforce consistency in both directions, thereby reducing instances of tracker losing objects and false positives. However, our MOTA and IDF1 metrics surpass those of QDTrack, with MOTA exceeding by 0.7 percentage points. This can be attributed to our use of multi-frame information during the data association stage, which enhances the accuracy and robustness of our tracking algorithm. Additionally, our trajectory association stage enhances the continuity and stability of MOT.

On the MOT20 test datasets, as shown in Table 3, our algorithm achieves the best performance in HOTA, IDF1, FP and IDS. Notably,

Table 4
DanceTrack test datasets. **Black bold** ranks first. SCTrack achieves rankings in HOTA, IDF1 and DetA.

Method	Ref.	MOTA↑	HOTA↑	IDF1↑	DetA↑	AssA↑
FairMOT [37]	IJCV 2021	82.2	39.7	40.8	66.7	23.8
ByteTrack [23]	ECCV2022	88.2	47.1	53.9	70.5	31.5
MOTR [39]	ECCV2022	79.7	54.2	51.5	73.5	**40.2**
GTR [48]	CVPR2022	84.7	48.0	50.3	72.5	31.9
OC-SORT [2]	CVPR2023	**92.0**	55.1	54.6	80.3	38.3
QDTrack [42]	TPAM2023	83.0	45.7	44.8	72.1	29.2
Baseline [29]	CVPR2023	91.3	56.7	57.7	81.1	39.8
SCTrack	–	91.7	**57.2**	**58.8**	**81.5**	40.1

while our MOTA metric did not reach the optimal level, this can be attributed to ByteTrack's use of a pure motion association method with lower memory consumption, leading to slightly higher overall MOTA performance. However, our algorithm significantly improves IDF1 by 0.9 compared to ByteTrack, demonstrating our substantial advantage in overall performance, encompassing object detection accuracy and tracking accuracy. Additionally, our IDS is reduced by 206 compared with other top-performing methods. This improvement is attributed we constructed, enhancing the reliability of data association. This further illustrates the superiority of our algorithm and affirm its capability to maintain excellent performance even in highly crowded and complex scenes.

On the DanceTrack test datasets, as shown in Table 4, our algorithm ranks first and second across all evaluation metrics. In contrast, the OC-SORT algorithm outperforms SCTrack in terms of MOTA, representing an improvement over kalman filter by utilizing non-linear motion prediction, particularly effective in scenarios with highly similar object appearances, reducing error accumulation. However, this approach also exhibits lower robustness under dim lighting conditions and heightened sensitivity to background interference. Conversely, our algorithm demonstrates significant improvements with increases of 2.1 in HOTA and 1.2 in IDF1, underscoring its advantages in handling high similarity scenarios. These results further demonstrate that our method not only

Table 5
Ablation experiment. Based on the baseline model, MOTA, HOTA, IDF1 and TP increase by 1.37%, 2.97%, 2.75% and 1447 respectively, while FN, FP and IDS decrease by 1509, 1575 and 29, indicating improved tracking performance and identity consistency.

Method	MOTA↑	HOTA↑	IDF1↑	FN↓	FP↓	TP↑	IDS↓
Baseline(BL)	76.20	66.86	78.43	13,940	7632	39,812	543
BL+MPAM	76.93	67.92	79.34	13,362	7096	40,390	529
BL+MPAM+MCDC	77.22	68.36	80.52	13,017	6696	40,735	525
BL+MPAM+MCDC+DCD	77.57(+**1.37**)	69.83(+**2.97**)	81.18+**2.75**)	12431(**–1509**)	6057(**–1575**)	41259(+**1447**)	514(**–29**)

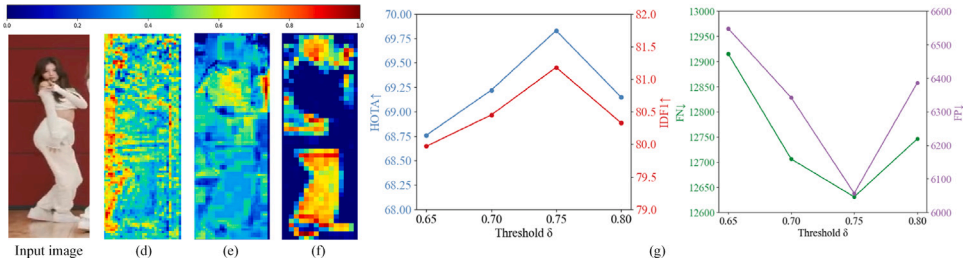

Fig. 7. Feature Map Visualizations. Fig. 7(d) shows the initial feature map, and Fig. 7(e) shows the rich semantic information extracted by F_4. Fig. 7(f) shows adding MDF-Fusion-Attention to obtain a more discriminative feature map. Fig. 7(g) shows trajectory association threshold δ. δ is 0.65, 0.7, 0.75, and 0.8 for experiments, we achieved the optimal value at 0.75.

maintains effectiveness, but also excels in tracking scenes with high similarity, which is crucial for addressing MOT challenges in practical real-world applications.

4.4. Ablation experiment

We conduct ablation experiments on the MOT17 datasets. In the experiment, the first half of the datasets are used as the training datasets, and the second half are used as the test datasets. The specific ablation experiments are shown in Table 5.

(1) Baseline (BL). The baseline model uses a pretrained model for feature extraction but does not train a ReID model, despite integrating a domain adaptation module. However, even with the addition of this module, the semantic information of the features is still not rich enough. In addition, the baseline model divides the tracks into active tracks and inactive tracks, but does not consider the associated matching of low-detection objects, so the overall algorithm is not robust. We innovate on these foundations and achieve significant improvements.

(2) BL+MPAM. By introducing the Multi-Perception Affinity Matrix (MPAM) module, the position, motion and appearance features are fused to construct a comprehensive multi-perception affinity matrix. This improvement improves the accuracy and robustness of the tracking algorithm, with MOTA increasing from 76.20 to 76.93, HOTA increasing from 66.86 to 67.92, and IDF1 increasing from 78.43 to 79.34. At the same time, the FN and FP are reduced to 13362 and 7096 respectively, and the number of TP increases to 40390. In addition, IDS decreased from 543 to 529, indicating a reduction in ID switches and more stable object identity maintenance. These results verify the effectiveness of the MPAM module in improving object tracking performance.

In addition, we also conducted a separate ablation experiment on Multi-Directional Feature Fusion Attention (MDF-Fusion Attention). As shown in Table 6, HOTA and IDF1 increased by 0.21 and 0.61 respectively, and FN and FP decreased by 649 and 405, indicating that the MDF-Fusion Attention module effectively enhances the directionality and attention mechanism of feature fusion, further improving the tracking accuracy. Fig. 7 shows the comparison of feature maps before and after adding the MDF-Fusion Attention module. Fig. 7(d) is the original feature map, Fig. 7(e) is the feature map containing rich context information extracted by $F4$, and Fig. 7(f) is the feature map after adding the MDF-Fusion Attention module. It can be clearly seen that the feature map after adding the MDF-Fusion Attention module is clearer and the features are more prominent.

Table 6
MDF-Fusion-Attention ablation experiment. After adding attention, the effect is significantly improved.

MDF-Fusion-Attention	HOTA↑	IDF1↑	FN↓	FP↓
–	67.71	78.73	14,011	7501
✓	67.92(+**0.21**)	79.34(+**0.61**)	13362(**–649**)	7096(**–405**)

(3) BL+MPAM+MCDC. On the basis of BL+MPAM, the Multi-frame Collaborative Distance Calculation (MCDC) module is further introduced. This method is for the association matching of low-detection objects. The introduction of this module improves MOTA to 77.22, HOTA to 68.36, IDF1 to 80.52, reduces FN to 13017, FP to 6696, and TP to 40735. IDS is further reduced to 525, demonstrating the effectiveness of multi-frame collaboration in maintaining object identity over time. This further illustrates the importance of the MCDC module for enhancing tracking performance and reducing false matches.

(4) BL+MPAM+MCDC+DCD. On the basis of the above modules, the Dynamic Confidence-Driven Trajectory Refinement (DCD) module is added. This module is a post-processing of the trajectory. Through this series of enhancements, MOTA is significantly improved to 77.57, HOTA reaches 69.83, and IDF1 reaches 81.18, which are all obvious improvements. At the same time, FN are significantly reduced by 586, FP are reduced by 639, and TP are increased by 524. IDS also dropped to 514, indicating that our post-processing module can further reduce identity switching and ensure consistency in long-term tracking. Overall, the trajectory breakage problem is effectively optimized and tracking loss is reduced, which further enhances the continuity and reliability of the overall tracking algorithm. Regarding the experiment of trajectory association threshold δ, we selected 0.65, 0.7, 0.75 and 0.8 for testing. The experimental results are shown in Fig. 7(g). It can be seen that when the threshold is set to 0.75, MOTA and IDF1 reach the highest value, while FP and FN are reduced the most.

In summary, after adding the three modules, compared with the baseline model, MOTA, HOTA and IDF1 increased by 1.37, 2.97 and 2.75 respectively, FN and FP decreased by 1509 and 1575 respectively, and TP increased by 1447. IDS was reduced from 543 to 514, showing a consistent decline across modules, and validating the stability of identity association throughout the tracking process. Through these improvements, the model showed significant improvements in various core indicators, verifying the effectiveness and robustness of the proposed method in improving object tracking performance. This series of optimization measures not only enhances the adaptability of the model in complex scenarios, but also improves the overall tracking accuracy and stability.

Fig. 8. MOT17 datasets visualization.

5. Visualization

Fig. 8 presents a visualization of the results obtained from our experiments on the MOT17 datasets. Scene 4 showcases the tracking performance under camera shake, demonstrating accurate object tracking despite the camera movement. Scenes 9 and 11 also exhibit excellent tracking performance when objects change scale and encounter frequent occlusions. These results highlight the method's capability to maintain high-precision tracking results in scenarios with variable object scales and frequent occlusions.

The experimental results on the MOT20 datasets are visualized in Fig. 9. As can be seen from scenarios 1, 2, 3, and 5, the scenario of MOT20 is more complex than that of MOT17, the algorithm performed well in the face of complex and dense crowds and maintained good tracking robustness. Despite the challenges posed by these scenes, the method effectively tracks objects, underscoring its strong performance in complex environments. These results further validate the reliability and robustness of the algorithm when dealing with dense crowds and complex scenes.

Fig. 10 visualizes the performance of the proposed method in crowded scenarios using the DanceTrack datasets. We carefully select four scenes designed to cover conditions in which DanceTrack can still maintain excellent tracking performance in low light, frequent motion, and similar appearances. These scenarios are selected not only with common challenges in mind but also to demonstrate robustness in real-world diverse environments.

6. Conclusion

In this paper, we introduces a Synergistic-aware Cascaded Association and Trajectory Refinement for Multi-Object Tracking (SCTrack). It mainly solves the problem of data association and trajectory management. SCTrack achieves significant performance improvements in MOT tasks, offering an effective solution to tackle object tracking challenges in complex scenes.

Compared with other methods, our method offers the following unique features and advantages:

(1) We created a synergistic-aware cascaded association module that integrates appearance, motion, and position features. This module incorporates MDF-Fusion attention to enhance feature extraction in three dimensions — horizontal, vertical, and across channels — to construct a robust MPAM. To handle low-confidence detections, we introduced multi-frame collaboration for association, constructed the MCDC module, and used historical information to guide the data association of the current frame.

(2) We address trajectory post-processing with a DCD module. This module enhances trajectory association and repairs trajectory fragments, particularly effective in scenarios with frequent occlusions, thereby improving trajectory continuity and integrity.

In particular, our method has demonstrated outstanding performance in practical applications. These modules enhance the model's adaptability in complex scenarios and substantially reduce false positives and false negatives. Experimental results show that our method has achieved significant improvements in multiple core indicators, confirming the effectiveness and robustness in object tracking tasks. The comprehensive integration of these innovative methods positions our tracking algorithm as superior to existing methods in terms of accuracy and stability.

Despite the strong performance of SCTrack, several limitations remain. The method currently relies heavily on the quality of object detection results, which may affect tracking robustness in low-quality detection scenarios. Additionally, the MCDC module introduces extra computational overhead, which may hinder real-time deployment in resource-constrained environments. To address these limitations, future research could explore optimizing computational efficiency through lightweight model design and efficient real-time deployment strategies

Fig. 9. MOT20 datasets visualization.

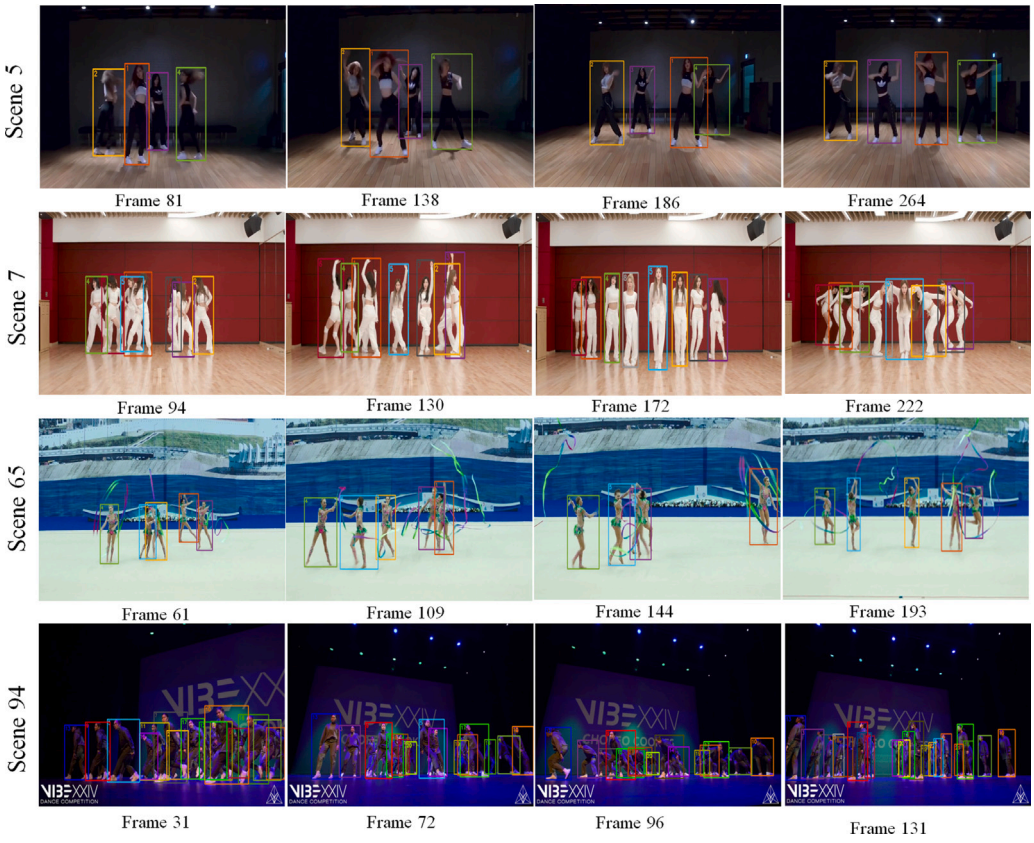

Fig. 10. DanceTrack datasets visualization.

on embedded platforms. These directions will further enhance the applicability of SCTrack in complex, real-world tracking scenarios.

CRediT authorship contribution statement

Hui Li: Methodology, Conceptualization. **Su Qin:** Writing – original draft. **Saiyu Li:** Visualization, Validation. **Ying Gao:** Writing – review & editing. **Yanli Wu:** Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in the National Key R&D Program of China (No. 2023YFF0612100), the Shandong Province Natural Science Foundation, China (No. ZR2024MF023), and the Key Technology Research and Industrial Demonstration Projects in Qingdao City, China (No. 23-7-2-qljh-4-gx, No. 24-1-2-qljh-19-gx).

Data availability

Data will be made available on request.

References

[1] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, P. Luo, Transtrack: Multiple object tracking with transformer, 2020, pp. 1–11, arXiv preprint arXiv:2012.15460.
[2] J. Cao, J. Pang, X. Weng, R. Khirodkar, K. Kitani, Observation-centric SORT: Rethinking sort for robust multi-object tracking, in: IEEE International Computer Vision and Pattern Recognition, 2023, pp. 9686–9696.
[3] Z. Qin, S. Zhou, L. Wang, J. Duan, G. Hua, W. Tang, MotionTrack: Learning robust short-term and long-term motions for multi-object tracking, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2023, pp. 17939–17948.
[4] N. Aharon, R. Orfaig, B.-Z. Bobrovsky, BOT-SORT: Robust associations multi-pedestrian tracking, 2022, pp. 1–13, arXiv preprint arXiv:2206.14651.
[5] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Fu, Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking, in: European Conference on Computer Vision, 2020, pp. 145–161.
[6] Y. Xiang, A. Alahi, S. Savarese, Learning to track: Online multi-object tracking by decision making, in: IEEE International Conference on Computer Vision, 2015, pp. 4705–4713.
[7] S.-H. Bae, K.-J. Yoon, Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2014, pp. 1218–1225.
[8] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J.T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al., A large-scale benchmark dataset for event recognition in surveillance video, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2011, pp. 3153–3160.
[9] L. Fei, B. Han, Multi-object multi-camera tracking based on deep learning for intelligent transportation: A review, Sensors 23 (8) (2023) 3852–3880.
[10] E.S. Candela, M.O. Pérez, C.M. Romero, D.C.P. López, G.S. Herranz, M. Contero, M.A. Raya, HumanTop: A multi-object tracking tabletop, Multimedia Tools Appl. 70 (2014) 1837–1868.
[11] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, T.-K. Kim, Multiple object tracking: A literature review, Artif. Intell. 293 (2021) 103448–103471.
[12] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, P. Luo, DanceTrack: Multi-object tracking in uniform appearance and diverse motion, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2022, pp. 20993–21002.
[13] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: Exceeding YOLO series in 2021, 2021, pp. 1–7, arXiv preprint arXiv:2107.08430.
[14] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multi-scale, deformable part model, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[15] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015) 1–9.
[16] F. Yang, W. Choi, Y. Lin, Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 2129–2137.
[17] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, CenterNet: Keypoint triplets for object detection, in: IEEE International Conference on Computer Vision, 2019, pp. 6569–6578.
[18] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[19] J. Redmon, A. Farhadi, YOLOV3: An incremental improvement, 2018, pp. 1–6, arXiv preprint arXiv:1804.02767.
[20] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, YOLOV4: Optimal speed and accuracy of object detection, 2020, pp. 1–17, arXiv preprint arXiv:2004.10934.
[21] D. Reis, J. Kupec, J. Hong, A. Daoudi, Real-time flying object detection with YOLOV8, 2023, pp. 1–10, arXiv preprint arXiv:2305.09972.
[22] C. Xiao, Q. Cao, Y. Zhong, L. Lan, X. Zhang, H. Cai, Z. Luo, D. Tao, MotionTrack: Learning motion predictor for multiple object tracking, 2023, arXiv preprint arXiv:2306.02585.
[23] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, X. Wang, ByteTrack: Multi-object tracking by associating every detection box, in: European Conference on Computer Vision, 2022, pp. 1–21.
[24] K. Huang, B. Sun, F. Chen, T. Zhang, J. Xie, J. Li, C.W. Twombly, Z. Wang, ReIDTrack: Multi-object track and segmentation without motion, 2023, pp. 1–8, arXiv preprint arXiv:2308.01622.
[25] Y.-H. Wang, J.-W. Hsieh, P.-Y. Chen, M.-C. Chang, H.H. So, X. Li, SMILEtrack: Similarity learning for occlusion-aware multiple object tracking, 2022, pp. 1–12, arXiv preprint arXiv:2211.08824.
[26] Z. Liu, X. Wang, C. Wang, W. Liu, X. Bai, SparseTrack: Multi-object tracking by performing scene decomposition based on pseudo-depth, 2023, pp. 1–13, arXiv preprint arXiv:2306.05238.
[27] G. Maggiolino, A. Ahmad, J. Cao, K. Kitani, Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification, 2023, pp. 1–5, arXiv preprint arXiv:2302.11813.
[28] Y. Jin, F. Gao, J. Yu, J. Wang, F. Shuang, Multi-object tracking: Decoupling features to solve the contradictory dilemma of feature requirements, IEEE Trans. Circuits Syst. Video Technol. (2023) 1–16.
[29] J. Seidenschwarz, G. Brasó, V.C. Serrano, I. Elezi, L. Leal-Taixé, Simple cues lead to a strong multi-object tracker, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2023, pp. 13813–13823.
[30] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, H. Meng, StrongSORT: Make deepsort great again, IEEE Trans. Multimed. (2023) 1–14.
[31] J. Wang, Y. Peng, X. Yang, T. Wang, Y. Zhang, SportsTrack: An innovative method for tracking athletes in sports scenes, 2022, pp. 1–7, arXiv preprint arXiv:2211.07173.
[32] G. Wang, Y. Wang, R. Gu, W. Hu, J.-N. Hwang, Split and connect: A universal tracklet booster for multi-object tracking, IEEE Trans. Multimed. 25 (2022) 1256–1268.
[33] J. Cao, Q. Chen, J. Guo, R. Shi, Attention-guided context feature pyramid network for object detection, 2020, pp. 1–12, arXiv preprint arXiv:2005.11475.
[34] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, J. Dong, Giaotracker: A comprehensive framework for MCMOT with global information and optimizing strategies in visdrone 2021, in: IEEE International Conference on Computer Vision, 2021, pp. 2809–2819.
[35] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, T. Mei, FastReid: A pytorch toolbox for general instance re-identification, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 9664–9667.
[36] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: The clear MOT metrics, EURASIP J. Image Video Process. 2008 (2008) 1–10.
[37] Y. Zhang, C. Wang, X. Wang, W. Zeng, W. Liu, FairMOT: On the fairness of detection and re-identification in multiple object tracking, Int. J. Comput. Vis. 129 (2021) 3069–3087.
[38] Q. Wang, Y. Zheng, P. Pan, Y. Xu, Multiple object tracking with correlation learning, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2021, pp. 3876–3886.
[39] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, Y. Wei, MOTR: End-to-end multiple object tracking with transformer, in: European Conference on Computer Vision, 2022, pp. 659–675.
[40] Z. Zhao, Z. Wu, Y. Zhuang, B. Li, J. Jia, Tracking objects as pixel-wise distributions, in: European Conference on Computer Vision, 2022, pp. 76–94.
[41] J. Cai, M. Xu, W. Li, Y. Xiong, W. Xia, Z. Tu, S. Soatto, MeMOT: Multi-object tracking with memory, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2022, pp. 8090–8100.
[42] T. Fischer, T.E. Huang, J. Pang, L. Qiu, H. Chen, T. Darrell, F. Yu, QDTrack: Quasi-dense similarity learning for appearance-only multiple object tracking, IEEE Trans. Pattern Anal. Mach. Intell. (2023) 1–18.
[43] J. Hyun, M. Kang, D. Wee, D.-Y. Yeung, Detection recovery in online multi-object tracking with sparse graph tracker, in: IEEE International Conference on Applications of Computer Vision, 2023, pp. 4850–4859.

[44] Y. Jian, C. Zhuang, W. He, K. Du, Y. Lu, H. Wang, Spatio-temporal correlation learning for multiple object tracking, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 6170–6174.

[45] L. Zheng, M. Tang, Y. Chen, G. Zhu, J. Wang, H. Lu, Improving multiple object tracking with single object tracking, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2021, pp. 2453–2462.

[46] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, W. Hu, Rethinking the competition between detection and REID in multi-object tracking, IEEE Trans. Image Process. 31 (2022) 3182–3196.

[47] D. Stadler, J. Beyerer, Modelling ambiguous assignments for multi-person tracking in crowds, in: IEEE International Conference on Applications of Computer Vision, 2022, pp. 133–142.

[48] X. Zhou, T. Yin, V. Koltun, P. Krähenbühl, Global tracking transformers, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2022, pp. 8771–8780.