A Novel Strategy for Multi-Objective Tracking Framework based on Semi-online Mechanism

Yanming Jin, Longjun Liu*, Hongbin Sun, Yizhuo Zhang, Nanning Zheng
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
*liulongjun@xjtu.edu.cn

***Abstract*—Multi-similar appearance and occlusion problem are still the main difficult issues in the field of multi-target tracking. Many practical multi-object tracking (MOT) algorithms based on detectors widely used in industry only rely on the detector's Intersection over Union (IOU) to solve most tracking problems. In this paper, we propose a MOT strategy based on a semi-online mechanism, which includes IOU matching strategy, trajectory exploration strategy and backtracking mechanism. Meanwhile, the integrated motion model and the enhanced appearance model are also presented. According to the tracking temporal window, our strategy will find multiple pairs of detection frames by matching the largest IOU first, and then it leverages Kalman filter to establish a motion model for exploring various tracking trajectory. At last, it calculates the reliability of different tracking trajectories and executes regret strategy we proposed to generate optimized tracking trajectory. Our algorithm focuses on transforming the tracking issues into how to generate robust fragments of tracking trajectory,and how to fix the back-tracking errors within the temporal window according to chronological order of tracking context. Our method is naturally suitable for solving the difficult issues such as disappearance, detection noise, and target occlusion. We evaluate our algorithms on the datasets of MOT2015 and MOT2017. Compared with baseline algorithm, our method reduces 16 IDS on each video sequence on average, and the multi-target tracking accuracy has been improved by 4.01%. Our algorithm model is very versatile and efficient for the algorithms of short trajectory stitching.**

Keywords—Semi-online, MOT, Regret Strategy, IOU Mask, Tracklet Association, Low IDS.

I. INTRODUCTION

Multi-object tracking is a fundamental problem in computer vision, which main calculates the trajectory of the target we concerned in a given video sequence. Traditional tracking algorithms can be roughly divided into online and offline tracking algorithms from the generation time of actual tracking results. Online tracking needs to complete real-time trajectory tracking immediately after each new frame detection operation is completed. Therefore, the online tracking algorithm intuitively has better real-time performance, but it cannot utilize global information of video frames and may lead to low accuracy. On the contrary, offline tracking is to track the trajectory after all frames of a given video sequence are detected. This mode can make good use of global information and has relatively more accurate tracking results, but it cannot meet the real-time requirements. In Multiple Object Tracking (MOT), all the current algorithms have a common feature: they are one-shoot results. Even if we have known that our tracking results have been wrong, it is difficult for us to make up for the errors. Therefore, in the past three years, a tracking algorithm called semi-online strategy has been proposed. We focus on the semi-online tracking mechanism in this study for improving the tracking accuracy and efficiency of MOT.

Current tracking algorithms tend to adopt complex appearance models to solve the similar issues in MOT, which use very complicated calculation methods to achieve more accurate appearance models, such as using convolutional neural network (CNN) to extract features, setting up RE-ID for tracking, etc. The complex algorithm may cause too much computing time overhead. In fact, many practical multi-object tracking algorithms based on detectors in industries only rely on the IOU of the detections to solve most disappearance or occlusion problems. Therefore, we believe that appearance model should not focus on how to completely recover all the appearance characteristics of the original object by appearance model, but the appearance model should reflect the discrimination between similar targets when involves the occlusions or similar issues. There are many reasons why the appearance model is difficult to distinguish, but most of them is because the IOU area between the target and the similar target is too large. Therefore, their appearance models are very similar in adjacent frames. To solve the problem, a very intuitive idea is to add mask to the IOU area, and the mask can increase the appearance discrimination in this case, which is one of the important issue we will study in this paper.

In this paper, we propose a novel strategy for multiple object tracking (MOT) based on semi-online mechanism, as shown in Figure 1, which is inspired by the study [13], our strategy also transforms MOT issue into matching problem among various small tracking trajectory in the temporal window (i.e., the Undirected Weighted Graph [13] in Figure 1), and takes full advantage of the global information of each fragment trajectory during a temporal window of 40 frames. Note, the disappearance problem and occlusion problem can hardly last for about 40 frames in most cases according to the conclusion of [13]. The scheme can effectively solve many occlusion problems and disappearance problems. But when the targets are dense enough to cause a large number of mutual occlusions, it will make the tracklet difficult to splice.

To solve the problem, we propose a new regret strategy to generate fragment tracklets. The regret strategy is based on Kalman Filter. We first look for a pair of Kalman heads in the temporal window. The length of temporal window is assumed as 20 frames and the start frame is set as ***Fs***. According to the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

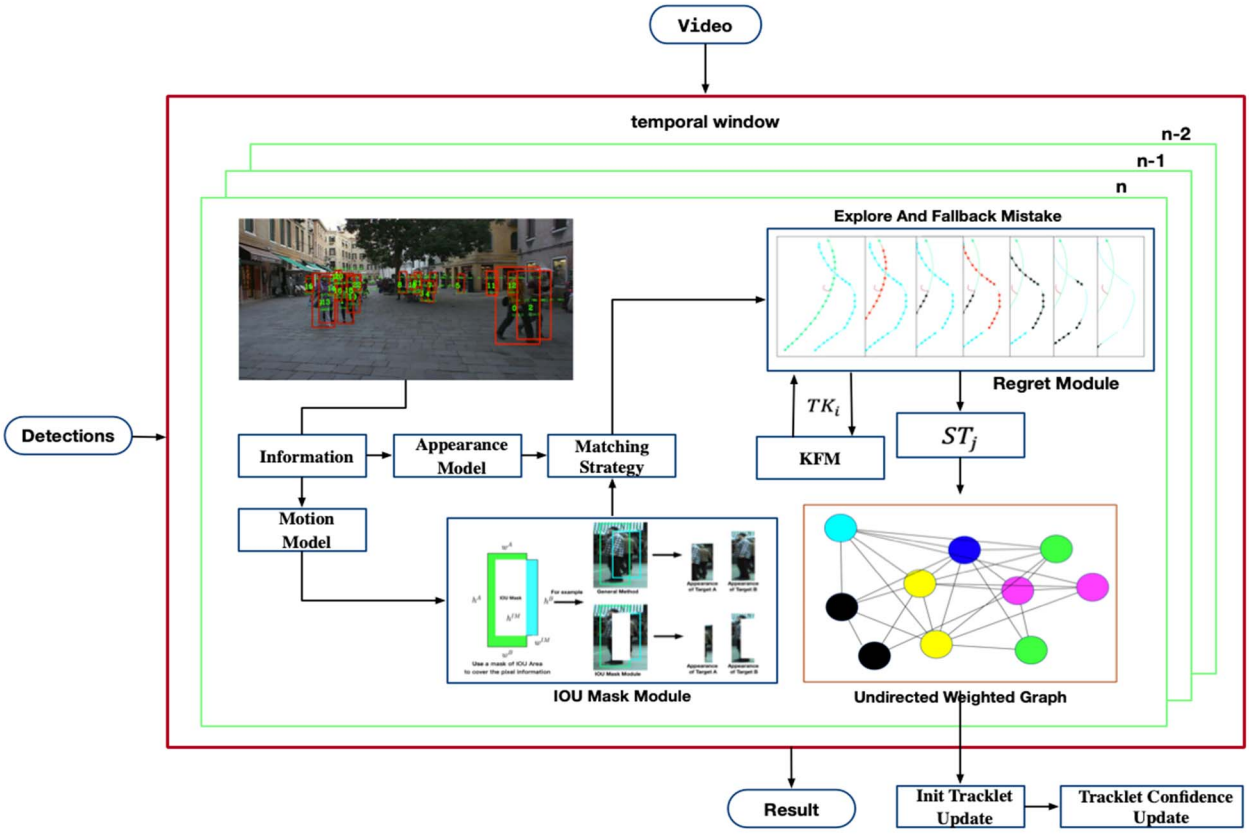

Figure. 1. The overview framework of multi-objective tracking (MOT) with our key modules

similarity between different IOU and the appearance model (pixel histogram with very little computation), our algorithm search a pair of Kalman heads $\boldsymbol{kh_1}$, $\boldsymbol{kh_2}$ starting from the window start position $\boldsymbol{Fs}$. Then we propose a motion model for the set of heads and predict the position $\boldsymbol{P}$ in next frame on the basis of the motion model. Suppose that the minimum length of the acceptable fragment trajectory is T_m, then we look for the next frame which has the most similar detection frame to $\boldsymbol{P}$, and the frame is spliced to the Kalman sequence meanwhile its motion model and appearance model will be updated. Through the iterations, in a subsequent frame, if the length of a Kalman sequence is found to be less than T_m, and we can't found any matching detection int the next frame, the subsequence will be disassembled. Otherwise, we track the Kalman sequence with a length longer than $\boldsymbol{N}$, making it a fragment track. Finally, the steps of splicing the growing trajectories of these fragment trajectories are performed. Our method has a good improvement over baseline in ETHMS, MOT15 and MOT17 in terms of IDS, MOTP and FP.

The main contributions of this article can be summarized as follows:

- In this paper, we propose a regret strategy that can explore various possible trajectories in the temporal window, and then generate a tracking trajectory. Finally, we implement the regret strategy and related motion model, appearance model and tracklet confidence for MOT.
- We present an IOU Mask strategy and matching strategy based on similarity to increase the discrimination of appearance model for the tracking targets that are obstructed largely with each other.
- The algorithm and strategy we proposed in this paper has good adaptability for current MOT tracking framework of fragment trajectory stitching based on semi-online mechanism.
- The computational overhead of our scheme is extremely low, the visual algorithms and graph calculations we used are very efficient and no precise Re-ID is required.

The remainder of this paper starts with a summary of background and related work in Section 2. Then we introduce our key modules and strategy for MOT framework in Section 3. We present our experimental results in Section 4. At last, we conclude our paper in Section 5

II. BACKGROUND AND RELATED WORK

This section will summarize online multi-object tracking, semi-online multi-object tracking and several typical methods related to this paper. As shown in Table 1, we list current the state-of-the-art MOT methods. These methods include Semi-

TABLE I. The comparison of current MOT methods

algorithm	Type	Avg Rank (MOT2015)	MOT	CNN
Semi MOT [13]	Semi-On-line	--	✓	×
CMOT [12]	On-line	--	✓	×
AlexTrac [18]	Off-line	61.7	✓	✓
Cem [17]	Off-line	49.5	✓	×
Tracktor15 [24]	On-line	34.6	✓	✓
KCF [25]	On-line	29.8	✓	✓
JointMC [26]	Off-line	28.1	✓	✓
AP_HWDPL_p [27]	On-line	22.1	✓	✓

Online methods, Online methods and Offline methods. The Avg Rank means the comprehensive performance rank from MOT2015 [22] challenge. Table1 also shows popular CNN based tracking frameworks, which is an emerging method with the development of deep learning technique.

A. Online Multi-target Tracking

In the past few years, many tracking algorithms [12, 14, 26, 27] including deep learning-based and detector-based tracking algorithms have been proposed. Algorithms of the two categories have their advantages and disadvantages. Among the tracking algorithms based on detector, [12] turns the MOT process into a multi-layer data association problem, opening up a classic idea. It proposes to transform the problem into three steps to improve performance: In the first step (low-level), by connecting detections of consecutive frames into short tracklets, and filtering unreliable short tracklets. In the second step (mid-level), the short tracklets obtained in the previous step are used to calculate the similarity, and the Hungarian algorithm [6] is used to stitch the short tracks to obtain longer tracks; In the third step (high-level) , refine the tracklets obtained in the mid-level. [12] proposes a robust online multi-target tracking algorithm, which proposes an idea of dimensioning the reliability of tracklets and learning online appearance models.

However, these algorithms have a common shortcoming——It is difficult to obtain relatively global information, the tracking still has relatively large problems when encountering difficult occlusion, similar problems, such as incorrect tracking results leading to the break of the tracking trajectory, causing a complete trajectory to be identified as two or more short trajectories.

B. Semi-Online Multiple Object Tracking

The paper [13] proposes a new tracking idea, which makes good use of global information in the process of exploring the relationship of each track during the tracking process by the temporal window. In fact, many traditional online multi-target tracking algorithms are difficult to effectively solve the occlusion problem for many reasons, e.g., the granularity of the prediction results of motion model is too large, and the distinction of appearance feature model is not obvious and so on. The design of the temporal window effectively combines the front and back information of the current frame. With future information and existed information of appearance model and motion models, it can effectively make up for the lack of information and information disorder caused by occlusion, disappearance and similar targets, etc. The short track splicing method with the temporal window is naturally suitable for solving the problem of occlusion,disappearance an so on. With the semi-online algorithm, the real trajectory is usually a splicing of several short tracks. Even if very severe occlusion or similarity issues occur, the problem of tracking error of the entire trajectory will be recovered and the tracking trajectory is still correct. Therefore, the Semi-online MOT[13] combined with context information has a very obvious effect on the occlusion problem. In this paper, we focus on the improvement for semi-online MOT framework.

III. Our Approach

In this section, we will present our key modules for MOT framework based on semi-online mechanism. Inspired by study [13], we present the overview framework of multi-objective tracking (MOT) with our key modules, as shown in Figure 1.

A. Motion model

We first define $\boldsymbol{X}$ and $\boldsymbol{Y}$ as any two tracklets, and the $\boldsymbol{v_X^F}$ and $\boldsymbol{v_Y^B}$ as the forward velocity from the head to the tail of $\boldsymbol{X}$, and the backward velocity from the tail to the head of $\boldsymbol{Y}$. We use $\mho$ to represent the process of motion simulated by Kalman Filter [4]. $\mathrm{F}(\boldsymbol{X},\boldsymbol{Y})$ is the forward affinity score obtained from the tail of $\boldsymbol{X}$ to the head of $\boldsymbol{Y}$. $\boldsymbol{\mathcal{B}}(\boldsymbol{X},\boldsymbol{Y})$ is the forward affinity score obtained from the head of $\boldsymbol{Y}$ to the tail of $\boldsymbol{X}$. Therefore, our motion affinity score can be defined as below:

$$\mathrm{F}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{\mathcal{L}}(\mho(\boldsymbol{p_X^{tail}} + \boldsymbol{v_X^F}), \boldsymbol{p_Y^{head}}) \quad (1)$$

$$\boldsymbol{\mathcal{B}}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{\mathcal{L}}(\boldsymbol{p_X^{tail}}, \mho(\boldsymbol{p_Y^{head}} + \boldsymbol{v_Y^B})) \quad (2)$$

$$\boldsymbol{\Lambda^M}(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{F} * \boldsymbol{\mathcal{B}} \quad (3)$$

Motion affinity score can be obtained by (3), where $\boldsymbol{\Lambda^M}(\boldsymbol{X},\boldsymbol{Y})$ represents the motion affinity between $\boldsymbol{X}$ and $\boldsymbol{Y}$ calculated by Kalman Fileter [4]. The bigger the value of $\boldsymbol{\Lambda^M}(\boldsymbol{X},\boldsymbol{Y})$ is, the more likely the $\boldsymbol{X}$ and $\boldsymbol{Y}$ are belong to the same target from the perspective of physical movement, which is the important clue for us to judge the relationship between two tracklets.

B. Appearance model

After conventional operations of cropping and resizing for the patch of detection in the n_{th} frame (i.e., we crop and resize each patch into a tensor, similarly to [64,128]), we can obtain that the number of patches equals to the number of detections. Then the pixels of $\boldsymbol{D_X^n}$ and $\boldsymbol{D_Y^n}$ will be grouped into several groups (i.e.,12 groups) by color interval (i.e., dividing the color into 64 intervals for every channel) and reshape the groups into one-dimension vector $\boldsymbol{Tsr_X}$ and $\boldsymbol{Tsr_Y}$ (i.e., we get a $\boldsymbol{1 \times 192}$

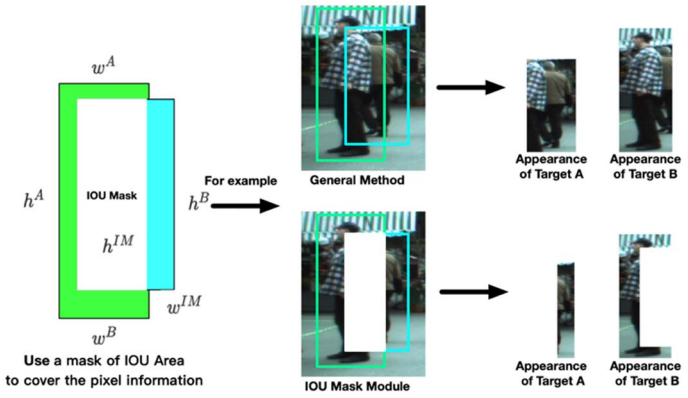

Figure. 2. The overview of our IOU Mask Module

tensor reshaped from $\mathbf{3 \times 64}$ which is obtained from RGB channels) for obtaining the appearance feature. According to the appearance models function in [12], we obtain the appearance model (i.e., we set our appearance model as a $\mathbf{1 \times 192}$ tensor): $\boldsymbol{f(X)}$ and $\boldsymbol{f(Y)}$. At last, we update the appearance model by vector fusion, which can be represent as $\mho(f(X), Tsr_X)$, $\mho(f(Y), Tsr_Y)$. Therefore, we can obtain the appearance affinity as shown below:

$$\mho(\boldsymbol{f(X)}, \boldsymbol{Tsr_X}) = \boldsymbol{\varphi f(X) + \theta Tsr_X} \quad \textbf{(4)}$$

$$\mho(\boldsymbol{f(Y)}, \boldsymbol{Tsr_Y}) = \boldsymbol{\varphi f(Y) + \theta Tsr_Y} \quad \textbf{(5)}$$

$$\boldsymbol{\Lambda^A(X,Y)} = \frac{\mho(f(X), Tsr_X) * \mho(f(Y), Tsr_Y)}{\|U^T f(X)\| \, \|U^T f(Y)\|} \quad \textbf{(6)}$$

The φ and θ is set to 0.9 and 0.1 respectively. $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be detections or tracklets. $\boldsymbol{\Lambda^A(X,Y)}$ is appearance model affinity, which is one of effective ways for us to deal with the situation that we can not figure out the complex relationship between tracklets (i.e., $\boldsymbol{TK_i}$ and $\boldsymbol{TK_j}$,) or detections (i.e., $\boldsymbol{D_X}$ and $\boldsymbol{D_Y}$) by physical motion simulation only. With the motion model and appearance model, we can better distinguish the stitching relationship of different tracklets.

C. Tracklet Confidence

Tracklet confidence can be intuitively understood as the matching degree between the constructed tracklet and the real track of the object. Here, we utilize the method base on [13] to represent the confidence of a tracklet ($conf(T_i)$) as shown below:

$$\boldsymbol{conf(T_i) = avg(z_k^i, z_j^i) * max((1 + \beta \cdot log\left(\frac{L-\alpha}{L}\right), 0)} \quad \textbf{(7)}$$

Where $\boldsymbol{avg(z_k^i, z_j^i)}$ denotes the average affinity between detections in an existing tracklets. $\boldsymbol{(z_k^i, z_j^i)}$ represents two detections $\boldsymbol{z_k^i}$ and $\boldsymbol{z_j^i}$ in tracklet $\boldsymbol{T_i}$ $\boldsymbol{max((1 + \beta \cdot log\left(\frac{L-\alpha}{L}\right), 0)}$ means the continuity of tracklet, $\boldsymbol{\alpha}$ is the number of frames where the object is missing. $\boldsymbol{\beta}$ is a control parameter related to the precision of the detector.

D. IOU Matching Strategy

Our similarity matching strategy is different from the similarity calculation of other typical MOT algorithms in [1] [2] [3] [13] [17]. We multiply different similarity factors to obtain a comprehensive similarity. We prioritize $\boldsymbol{\Lambda^{IOU}}$ as the first screening factor. The mathematical formula can be expressed as follows

$$\boldsymbol{\Lambda^{IOU}(D_1, D_2) = \frac{D_1 \cap D_2}{D_1 + D_2 - (D_1 \cap D_2)}} \quad \textbf{(8)}$$

$$\boldsymbol{\Lambda^{IOU}(D_1, D_2) > thres^{IOU}} \quad \textbf{(9)}$$

$$\boldsymbol{\Lambda^{A}(D_1, D_2) > thres^{A}} \quad \textbf{(10)}$$

$$\boldsymbol{\Lambda^{match}(D_1, D_2) = (7) \cap (8)} \quad \textbf{(11)}$$

When we encounter multiple candidates that meet the IOU threshold: $\boldsymbol{thres^{IOU}}$, we will further add the appearance and size similarity to target discrimination factors. When the objects are difficult to distinguish from the target, $\boldsymbol{\Lambda^{match}}$ will be sorted to choose the most similar candidate detection.

E. IOU Mask Module

We design the IOU mask module to deal with the scenarios that two or more targets are occluding each other. Figure 2 shows that the scenario of mutual occlusion. When targets **A** and **B** are occluded from each other, we do not directly extract features from the location of detection, but use the IOU area between **A** and **B** as an IOU mask to cover the pixel information of the IOU area when extracting the appearance features. Our mask strategy can efficiently increase the discrimination of different targets. Unfortunately, when multiple targets are blocking each other, it is very easy lead to the detection area of a target to be almost completely covered by multiple IOU masks. To avoid situation like this, we set IOU mask threshold to avoid the worst case that we can't get any appearance information.

We formulize the IOU mask operation as follows:

In the n_{th} frame, we have multiple detections: $\boldsymbol{D^n}$, whose area is $\boldsymbol{S^n}$. The set of IOU masks among them is $\boldsymbol{L^{Mask}}$. For the $\boldsymbol{k_{th}}$ detection, suppose there are $\boldsymbol{M_k}$ IOU Masks covering area of $\boldsymbol{D_k^n}$, then we define these IOU masks set as L_A^{Mask}. If the remaining area of $\boldsymbol{D_k^n}$ after being covered by $\boldsymbol{L_A^{Mask}}$ is $\boldsymbol{AR_A}$, the coverage process can be expressed as $\boldsymbol{D_k^n \otimes L_A^{Mask}}$:

$$\boldsymbol{AR_A = \frac{D_k^n \otimes L_A^{Mask}}{S_k^n}} \quad \textbf{(12)}$$

in case

$$\boldsymbol{AR_A < Thres^{IM}} \quad \textbf{(13)}$$

Then, we remove the IOU mask whose area is the smallest in $\boldsymbol{L_A^{Mask}}$ and recalculate equation 12 until AR_A meets the expression of equation 13.

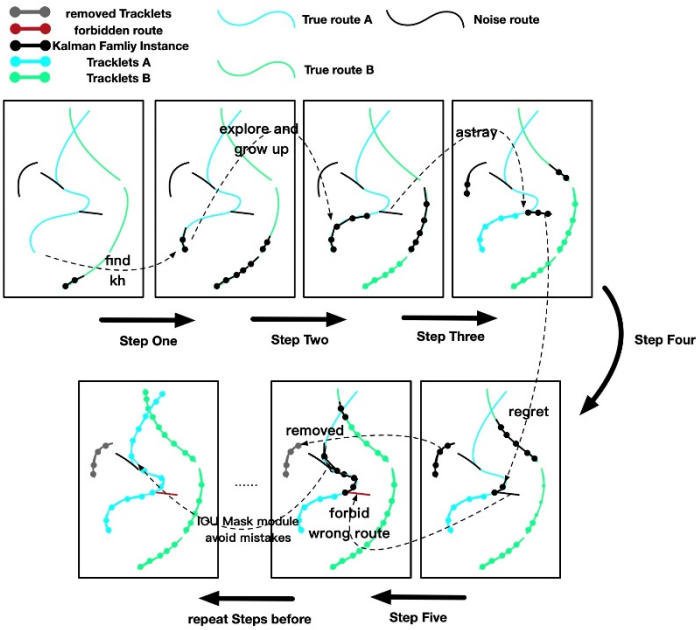

Figure 3: Regret strategy is able to achieve corrected trajectory in the temporal window, while finding some wrong routes during trajectory exploration.

F. Regret Strategy for Semi-Online MOT

In this section we will introduce our regret strategy based on semi-online multi-target tracking framework, as shown in Figure 1 (in the top right corner).

We define the length of the temporal window is $\boldsymbol{N}$. The minimum instantiation length of the short tracklets is $\boldsymbol{T_m}$. The Kalman family map can be denoted as $\boldsymbol{KFM}$, which is used to record the relationship between the detections by motion model and appearance model (i.e.,$\boldsymbol{KFM}$ in Figure 1). Note that the k_{th} detections in the n_{th} frame is denoted as D_k^n, which also contains the coordinates and reliability of a detection in a List: [**x, y, w, h, conf**]; We use $\boldsymbol{KFR_k^n}$ denote the List of detection D_k^n in the $\boldsymbol{KFM}$. It can be represented by the following mathematical expression:

$$KFR_k^n = x,\ x \epsilon N \qquad \textbf{(14)}$$

If $\boldsymbol{KFR_k^n} = -1$, it represents the detection has nothing to do with the $\boldsymbol{KFM}$. The $\boldsymbol{x}$ means the detection for the $(\boldsymbol{x}+\boldsymbol{1})_{\boldsymbol{th}}$ member of a short tracklet in the exploration state in $\boldsymbol{KFM}$. The i_{th} short tracklets in $\boldsymbol{KFM}$ is defined as $\boldsymbol{TK_i}$, which is longer than $\boldsymbol{T_m}$ and not updated in the x_{th} frame. It will be instantiated as reliable tracklets $\boldsymbol{ST}_j$,which are carefully selected by regret strategy.

Next, we will present the process of short tracklets exploration and regret strategy for trajectories at the $\boldsymbol{n_{th}}$ frame in detail, as shown in Figure 3:

Step One, ***Find kh***: Among the detections $(\boldsymbol{D_i^n}, \boldsymbol{D_j^{n+1}})$, in the n_{th} frame and $(n+1)_{th}$ frame we look for every pair of detections that may belong to the same tracklets based on IOU. If $\boldsymbol{KFR_i^n} = -1$ and $\boldsymbol{KFR_j^{n+1}} = -1$,we will mark $\boldsymbol{KFR_i^n}$and $\boldsymbol{KFR_j^{n+1}}$as 0 and 1. Respectively, we will temporarily call every pair of detections as $\boldsymbol{kh}$. After this step, we will have several pairs of $\boldsymbol{kh}$ (i.e., $\boldsymbol{KFR_i^n}$and $\boldsymbol{KFR_j^{n+1}}$) in n_{th} frame and $(n+1)_{th}$ frame.

Step Two, ***Predict and Explore*****:** We have obtained an unstable tracklet, $\boldsymbol{TK}_i$, which is composition of some $\boldsymbol{kh}$ (i.e., $\boldsymbol{KFR_i^n}$and $\boldsymbol{KFR_j^{n+1}}$). We then build motion model for each unstable tracklet that has not yet been instantiated with $\boldsymbol{TK_i}$. Based on the motion model, we predict the position of the target $\boldsymbol{TK_i}$ at $(n+2)_{th}$ frame, and we define the predicted position of $\boldsymbol{TK_i}$at $(n+2)_{th}$ frame as $\boldsymbol{PP_i^{n+2}}$. Note, the position prediction method is introduced in [12].

Step Three, ***Grow up*****:** According to matching strategy mentioned in section 3.D, we select detection $\boldsymbol{D_k^{n+2}}$ that is the most similarity position to $\boldsymbol{PP_i^{n+2}}$ and mark the $\boldsymbol{KFR_k^{n+2}}$ in the $\boldsymbol{KFM}$ as $\boldsymbol{KFR_j^{n+1}}+1$. Then we update the motion and appearance information of $\boldsymbol{TK_i}$. After that, we use the updated motion model to predict the position in the next frame through method of Step Two.

Step Four: Repeat Step One, Step Two and Step Three.

Step Five, ***Instantiation or Regret*****:** After performing the steps above, we make an instantiation or regret decision on short tracklets in $\boldsymbol{KFM}$ (i.e., $\boldsymbol{TK_0}, \boldsymbol{TK_1}, \ldots, \boldsymbol{TK_i}$) based on the following conditions：

a) If the length of the Kalman Tracklet such as $\boldsymbol{TK}_i$ exceeds or equals the threshold $\boldsymbol{T_m}$ and it is not updated in the last frame ,we will instantiate $\boldsymbol{TK}_i$ as new reliable tracklets: $\boldsymbol{ST}_j$.

b) If the length of the Kalman Tracklet such as $\boldsymbol{TK}_i$ is less than T_m,and it does not be updated in the last frame, we will disassemble $\boldsymbol{TK}_i$ in $\boldsymbol{KFM}$ and set all the $\boldsymbol{KFR}$ which belong to $\boldsymbol{TK}_i$ to -1, and mark this track as a forbidden route to avoid it is explored in next steps.

After the five steps above, we will obtain reliable short tracklets as shown in Figure 1 and Figure 3.

IV. Experiments

In this section, we will demostrate our experimental results on MOT2015 [22], MOT20179 [23] and ETHMS BANHNHOF [22] datasets, and compare the results with other typical tracking algorithms.

A. Experimental Methodology

We implement the Semi-online MOT with our key strategies by python3 on a PC with a CPU of 2.2Ghz and 32G memory. And all parameters are adjusted based on our various experiment results. We set the IOU MASK threshold as 0.2, the overall similarity threshold is set as 0.85, and the length of our temporal window to 40 frames according to the result of [13].

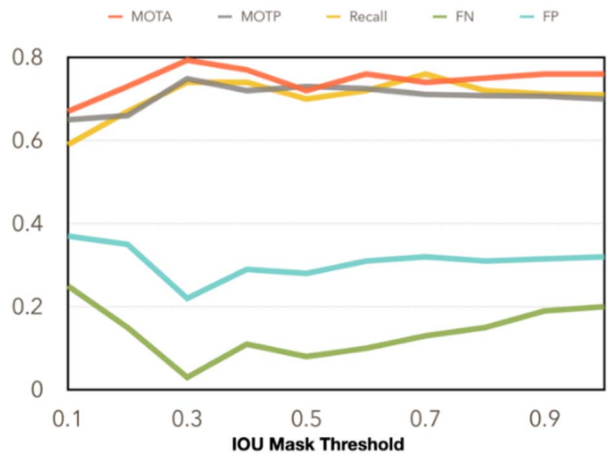

Figure. 4. The performance achieved by the proposed MOT method under different settings of IOU Mask Threshold

B. Evaluation Indicators

Note that, the measures results denoted by (↑) means a higher score, which indicates better performance. On the contrary, the measure results denoted by (↓) means the worse performance.

Similar to [13], we used the number of all ground truth Trajectories (GT), the number of Mostly Tracked trajectories (MT), the number of Mostly Lost trajectories (ML), Recall (↑), Precision (↑), MOTA (↑) and MOTP (↑) for performance evaluation. MOTA (↑) and MOTP (↑) are used to compare the performance of different MOT algorithms. MOTA combines three different errors, i.e., False Negatives (FN ↓), False Positives (FP ↓), and Identity Switching (IDS ↓) [13]. MOTP is calculated as the overlap between the ground truth tracks and the estimated tracks. It shows the ability of a tracker to estimate the precise object positions[19].

C. IOU Mask Threshold Settings

We tested the performance of the proposed method under different settings of IOU Mask Threshold. Figure. 4 shows the experiment result on the ETHMS BAHNHOF dataset.

It is obvious that our method has a best performance while IOU Mask Threshold is about 0.3. When the IOU Mask grows to 1, our method performs worse and the performance is almost the same as our method without IOU Mask module. This is because the IOU Mask Module is able to increase the appearance discrimination when the target is heavily obscured. The larger the IOU Mask Threshold means that the IOU Mask module can only be used when the target has less occlusion or no occlusion. When the IOU Mask threshold is less than 0.3 and continues to decrease, various performance indicators such as MOTA, MOTP and other indicators have decreased by about 12% and 8%, respectively. Because an IOU Mask Threshold that is too small will make the target almost completely lose appearance information under the influence of the IOU Mask, resulting in a situation where the appearance model is temporarily disabled. In summary, when the IOU Mask Threshold is about 0.3, the IOU Mask Module has the effect of improving the appearance discrimination of a target which is severely blocked, thereby improving various indicators.

D. Comparison of Experimental Results

In this section, we will analyze the experimental results of our method on five classic MOT datasets from various indicators. To be fair, we used the same detections and GTs, and all experiments were performed on the same data set, as shown in Table 2 (Next page).

ETHMS BAHNHOF: The difficulties in this video sequence include a small amount of image interference, a small amount of noise detection and mutual occlusion. The difficulty is relatively low overall. According to our experimental data, our method performs better than baseline [13] in MOTA, MOTP, FP, IDS, ML and other indicators, especially MOTA and MOTP are 2.37% and 4.58% higher respectively. And we get lower FP, higher Recall and MT. Our method shows stable and strong overall performance on most datasets.

MOT15 (TUD-Stadtmitte): This video sequence has the difficulty of low viewing angle, a large number of severe occlusions with each other and complete occlusion. The scenario is more challenging. However, we can see that our method still obtains the best results on multiple indicators such as MOTA, MOTP, FP, IDS, especially MOTA is 4.94% higher than the baseline [13]. And FP, FN, MT, ML and other indicators have proved that our method is extremely small and stable in terms of Error rate which is also reflected in subsequent video sequences. Most important indicators of MOT15 (TUD-Campus) also exceed the baseline [13], and the trend is similar.

The main challenge in PETS (S2L1) is that the target moves in high-speed non-linear mode, and the target is frequently blocked. PETS (S2L2) has more severe occlusion. We can see from the experimental data that CMOT[12], CEM[17] and baseline[13] have obtained the best performance of each index in PETS (S2L1) and PETS (S2L2). The various indicators of our method have obtained a relatively high ranking in these two data sets. Although not prominent, our method has a very stable performance. The difference between MOTA and MOTP of CMOT [12] in PETS (S2L2) reached 16.2%; The difference between MOTA and MOTP of baseline[13] in PETS (S2L1) reached 11.38%. The difference of our methods between MOTA and MOTP is around 3.5% which is very stable and very close to the best performance. However, their performance is very unstable on different data sets. But our method has good performance and stability among various indicators and different data sets.

V. Conclusion

In this paper, we propose a retrospective multi-target tracking strategy based on semi-online MOT, which includes IOU matching strategy, trajectory exploration strategy and backtracking mechanism. Meanwhile, the motion model and the enhanced appearance model are presented to explore the motion trajectory, proactively search the wrong trajectory, and generate a series short trajectory for high tracking accuracy and tracking

TABLE II. Performance comparison of different MOT algorithms with the strategy we proposed. The best performance is marked with bold for each dataset.

Datasets	algorithm	MOTA ↑	MOTP ↑	FP ↓	FN ↓	IDS ↓	Rcll ↑	GT -	MT ↑	ML ↓
ETHMS BAHNHOF	Baseline [13]	76.96%	70.29%	32.19%	**2.28%**	16	71.27%	126	96	**10**
	Proposed	**79.33%**	**74.87%**	**22.04 %**	2.99%	**13**	**74.29 %**	126	**104**	13
MOT15 (TUD-Campus)	CMOT [12]	51.23%	70.46%	36.58%	39.58%	3	-	8	-	-
	Alextrac [18]	35.7%	65.6%	40.67%	13.21%	23	-	8	1	**1**
	Baseline [13]	62.58%	71.42%	29.36%	**12.82%**	**3**	63.09%	8	**2**	**1**
	Proposed	**64.25%**	**72.37%**	**28.81 %**	14.31%	4	64.11%	8	**2**	**1**
MOT15 (TUD-Stadtmitte)	CMOT [12]	45.23%	70.46%	27.48%	18.63%	16	-	9	-	-
	Alextrac [18]	53.8%	65.6%	24.8%	**10.42%**	40	**84.70 %**	9	5	**0**
	Baseline [13]	67.42%	71.64%	12.56%	15.32%	6	79.65%	9	**7**	**0**
	Proposed	**72.36%**	**73.92%**	**12.01 %**	14.55%	5	83.51%	9	**7**	**0**
PETS (S2L1)	CMOT [12]	83.04%	69.59%	**1.19%**	19.41%	**4**	-	23	**23**	**0**
	CEM [17]	80.20%	**90.60%**	12.15%	**2.37%**	11	-	23	21	1
	Baseline [13]	**93.35%**	81.97%	4.25%	3.91%	5	95.61%	23	**23**	**0**
	Proposed	90.20%	87.34%	3.47%	8.01%	**4**	93.22%	23	**23**	**0**
PETS (S2L2)	CMOT [12]	**70.12%**	53.92%	**14.99 %**	14.35%	**45**	-	74	53	**1**
	CEM [17]	56.90%	59.40%	27.90%	6.04%	99	**65.50 %**	74	28	12
	Baseline [13]	67.21%	**71.30%**	33.04%	**1.83%**	60	59.67%	74	60	8
	Proposed	65.43%	69.07%	26.70%	4.71%	51	63.01%	74	**66**	6

efficiency. Compared with baseline[13], our algorithm can save a lot of computing resources (about 28.9%) on MOT, and exceed or approach the baseline[13] on various indicators. Moreover, our algorithm has a good universal applicability to MOT with short track stitching. Our strategy has been evaluated on multiple public datasets, and experimental data shows that our algorithm has a significant improvement over existing studies.

ACKNOWLEDGMENT

This research was supported by National Natural Science Foundation of China (No. 61722406, 61751401, 61602368), and by the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Tang, S., Andriluka, M., Andres, B., and Schiele, B.,"Multiple people tracking by lifted multicut and person re-identification," Proc. IEEE Conf. Comput. Vis. Pattern Recognit,pp. 3539-3548,2017.

[2] Tang, S., Andres, B., Andriluka, M., and Schiele, B., "Subgraph decomposition for multi-target tracking," Proc. IEEE Conf. Comput. Vis. Pattern Recognit,pp. 5033-5041,2015.

[3] Yang, B., and Nevatia, R., "Multi-target tracking by online learning a CRF model of appearance and motion patterns," International Journal of Computer Vision, vol.107(2), pp.203-217, 2014.

[4] Bishop G, and Welch G. ,"An introduction to the kalman filter[J]," Proc of SIGGRAPH, Course, vol.8,pp.27599-23175,2001.

[5] Cuevas, E. V., Zaldivar, D., and Rojas, R., "Kalman filter for vision tracking",2005.

[6] Ahuja, Ravindra K. , R. K. Ahuja , and R. K. Ahuja ,"Network Flows. Optimization. Elsevier North-Holland", 1989.

[7] Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B., "Simple online and realtime tracking," IEEE International Conference on Image Processing (ICIP) ,pp. 3464-3468, September 2016.

[8] Wang, B., Wang, G., Chan, K. L., and Wang, L., "Tracklet association by online target-specific metric learning and coherent dynamics estimation," IEEE Trans. Pattern Anal. Mach. Intell, vol. 39,no.3,pp. 589-602, 2016.

[9] Xiang, Y., Alahi, A., and Savarese, S., "Learning to track: Online multi-object tracking by decision making," In Proceedings of the IEEE international conference on computer vision ,pp. 4705-4713 ,2015.

[10] Huang, C., Wu, B., and Nevatia, R., "Robust object tracking by hierarchical association of detection responses," In European Conference on Computer Vision ,pp. 788-801,October 2008.

[11] Li, Y., Huang, C., and Nevatia, R., "Learning to associate: Hybridboosted multi-target tracker for crowded scene," Proc. IEEE Conf. Comput. Vis. Pattern Recognit ,pp. 2953-2960,June 2009.

[12] Bae, S. H., and Yoon, K. J., "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning,"Proc. IEEE Conf. Comput. Vis. Pattern Recognit,pp. 1218-1225,2009.

[13] Wang, J., Guo, Y., Tang, X., Hu, Q., and An, W., " Semi-online multiple object tracking using graphical tracklet association," IEEE Signal Processing Letters, vol.25,no.11,pp.1725-1729,2018.

[14] Dalal, N., and Triggs, B., "Histograms of oriented gradients for human detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit,vol. 1, pp. 886-893,June 2005

[15] Fisher, R. A., "The use of multiple measurements in taxonomic problems," Annals of eugenics, vol.7,no.2,pp. 179-188,1936.

[16] Hu, Q., Guo, Y., Lin, Z., An, W., and Cheng, H., "Object tracking using multiple features and adaptive model updating," IEEE Transactions on Instrumentation and Measurement, vol.66,no.11, pp.2882-2897,2017.

[17] Milan, A., Roth, S., and Schindler, K. , "Continuous energy minimization for multitarget tracking," IEEE transactions on pattern analysis and machine intelligence, vol.36,no.1, pp.58-72,2013.

[18] Bewley, A., Ott, L., Ramos, F., and Upcroft, B., "Alextrac: Affinity learning by exploring temporal reinforcement within association chains," In 2016 IEEE International conference on robotics and automation (ICRA) ,pp. 2212-2218, May 2016.

[19] Bernardin, K., and Stiefelhagen, R., "Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP Journal on Image and Video Processing," pp. 1-10,2008.

[20] Wang, C., Liu, H., and Gao, Y.,"Scene-adaptive hierarchical data association for multiple objects tracking," IEEE Signal Processing Letters, vol 21,no. 6, pp.697-701,2014.

[21] Yi, Y., and Xu, H.,"Hierarchical data association framework with occlusion handling for multiple targets tracking," IEEE signal processing letters, vol. 21,no. 3, pp.288-291,2014.

[22] Leal-Taixé, Laura, "MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking," 2015.

[23] Dendorfer, Patrick, "CVPR19 Tracking and Detection Challenge: How crowded can it get?," 2019.

[24] Bergmann, P. , Meinhardt, T. , and Leal-Taixe, L., "Tracking without bells and whistles,"2019.

[25] P. Chu, H. Fan, C. Tan, H. Ling.,"Online Multi-Object Tracking with Instance-Aware Tracker and Dynamic Model Refreshment,"In WACV, 2019.

[26] Keuper, Margret, Tang, Siyu, Andres, Bjorn, Brox, Thomas, and Schiele, Bernt., "Motion segmentation & multiple object tracking by correlation co-clustering," IEEE Trans. Pattern Anal. Mach. Intell, pp.1-1,2018.

[27] C. Long, A. Haizhou, S. Chong, Z. Zijie,and B. Bo., "Online Multi-Object Tracking with Convolutional Neural Networks,"In 2017 IEEE International Conference on Image Processing (ICIP), 2017.