Group Guided Data Association for Multiple Object Tracking

Yubin Wu[1,2][0000−0002−8504−7662], Hao Sheng[1,2,3][0000−0002−2811−8962], Shuai Wang[1,2][0000−0002−1570−6570], Yang Liu[1,2][0000−0001−9698−9551], Zhang Xiong[1,2,3], and Wei Ke[3]

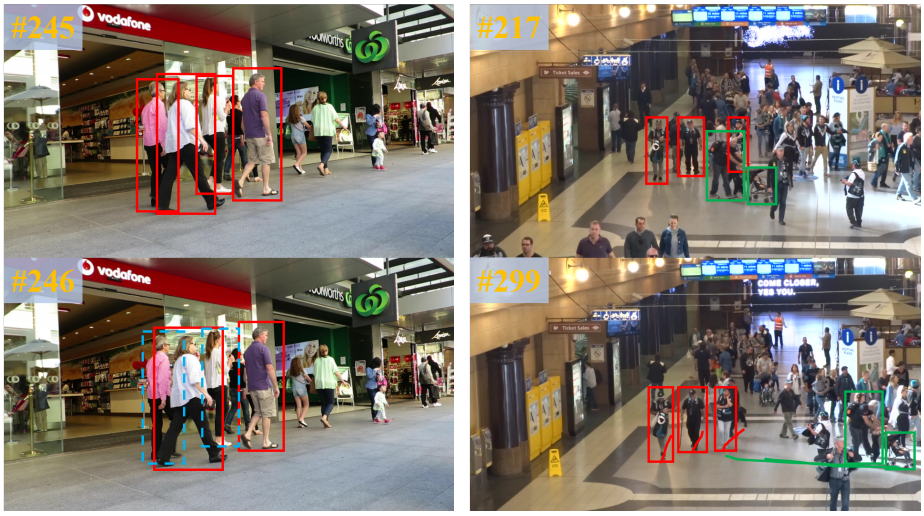
[1] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China
[2] Beihang Hangzhou Innovation Institute Yuhang, Xixi Octagon City, Yuhang District, Hangzhou, China
[3] Faculty of Applied Sciences, Macao Ploytechnic University, Macao SAR, China
{yubin.wu, shenghao, shuaiwang, liu.yang, xiongz}@buaa.edu.cn, wke@mpu.edu.mo

Abstract. Multiple Object Tracking (MOT) usually adopts the Tracking-by-Detection paradigm, which transforms the problem into data association. However, these methods are restricted by detector performance, especially in dense scenes. In this paper, we propose a novel group-guided data association, which improves the robustness of MOT to error detections and increases tracking accuracy in occlusion areas. The tracklets are firstly clustered into groups of related motion patterns by a graph neural network. Using the idea of grouping, the data association is divided into two stages: intra-group and inter-group. For the intra-group, based on the structural relationship between objects, detections are recovered and associated by min-cost network flow. For inter-group, the tracklets are associated with the proposed hypotheses to solve long-term occlusion and reduce false positives. The experiments on the MOTChallenge benchmark prove our method's effects, which achieves competitive results over state-of-the-art methods.

Keywords: Multiple Object Tracking · Target Grouping · Data Association.

1 Introduction

Multi-Object Tracking (MOT) is an essential topic in computer vision and is widely used in video understanding, intelligent transportation[30], and surveillance systems. Benefiting from the progress of detectors, MOT methods usually adopt the Tracking-by-Detection paradigm, which associates detections with object identities. However, in the case of frequent object interaction and dense occlusion in practical applications, detetors are often performed with errors. Consequently, it is difficult for trackers to recover the missed detection; on the other hand, appearance metrics are no longer reliable when objects overlapped. These problem become the main challenge for the MOT methods.



(a) A group with missed detections (b) Two group intersect

Fig. 1. The examples of object groups, where frame numbers are colored in yellow. (a) The red boxes denotes detections while the blue dashed boxes denotes missed detections. (b) The red and green boxes represent two different groups of objects respectively.

According to the different needs of online[38, 29] and batch processing[36, 35], association algorithms are generally divided into two types: the one is local optimization, such as Bipartite Graph[3], and Heuristic Hypothesis[8]. The other is global optimization, such as Network Flow[40], and Conditional Random Field[21]. The local optimization algorithms are more robust to cumulative detection errors but at the expense of the ability to handle long trajectories. The achievements of these methods are mainly attributed to recovering missed detections by regression and searching with additional detectors. The global optimization algorithms maintain higher trajectory integrity but are disturbed by detection errors due to delayed decisions. These methods focus on feature modeling of long trajectories for the anti-occlusion association.

Motivation: Aiming at the above problems, we designed a novel data association method with group information. As the main target of MOT, pedestrians often move as different groups by companions and roads. As shown in Fig. 1 (a), a group of pedestrians have a similar movement pattern and maintain a stable relative position for a short time. In addition to entering and leaving the scene, the number of pedestrians in the group also remained stable. This inspired us to recover the missed detections (blue dashed boxes in Fig. 1 (a)) by position and number constraints when occlusion occurs. As shown in Fig. 1 (b), pedestrians in different groups continue to move independently after a short occlusion. Therefore, associating pedestrians with groups respectively can avoid the interference

caused by occlusion. In summary, this requires a data association method with group granularity.

In this paper, we propose a data association method guided by group information for better tracking accuracy in dense scenes. In the MOT problem, objects association can be naturally transformed into a graph problem. Based on previous work[34], we design a graph network to cluster objects into groups. For intra-group association, the method assumes the number and relative position of objects are maintained stable. We construct dummy detections for the intra-group association to meet the constraints and use the min-cost network flow model to obtain the tracklets. Although the dummy nodes recover missing detection, it brings some false positives in tracklets. Therefore, in the inter-group association, we establish the association hypotheses for tracklets. Before solving the hypotheses, the pruning strategy is used to reduce the false positives, and then the algorithm measure appearance affinity to get complete trajectories. Experiments show that our method significant improvement in detection recovery and long-term association.

In summary, our main contributions include:

- Design a graph network to obtain tracklets group by aggregating objects motion information.
- Propose Intra-Group association by min-cost network flow, which recover missing detection by constraint in group.
- Propose Inter-Group association by hypotheses proposal of tracklets, which use pruning to reduce false positives and measure appearance affinity to solve long-term occlusion problem.

2 Related Work

The data association problem in multi-target tracking aims to distinguish multiple identity tags of detections. In association measurement, one kind of method can achieve better tracking by adopting multi-stage[13] and multi-granularity association strategy[27]. The other method uses multiple features[28] to constrain the feasible solution space in each decision window[22, 41], so as to reduce the computational overhead and difficulty. We consider combining the advantages of these two types of methods. The proposed method uses group-guided two-stage association from different granularity and distinguishes different occluded targets, which reduces the interference in the calculation. To achieve higher accuracy, grouping and two-step association rely on global information, so our method is batched and not designed for real-time systems.

Motion pattern analysis based on social groups contributes to improving the accuracy of MOT in dense scenes. Pellegrini. et al[23] proposed data association by joint modeling of pedestrian trajectories and groupings. Zhao. et al[44] proposed a tracking method using motion patterns for very crowded scenes. Kratz. et al[16] proposed a tracking method using local spatio-temporal motion patterns in extremely crowded scenes. These methods mainly study the use of groups to predict the future movement of targets and the structural information of the

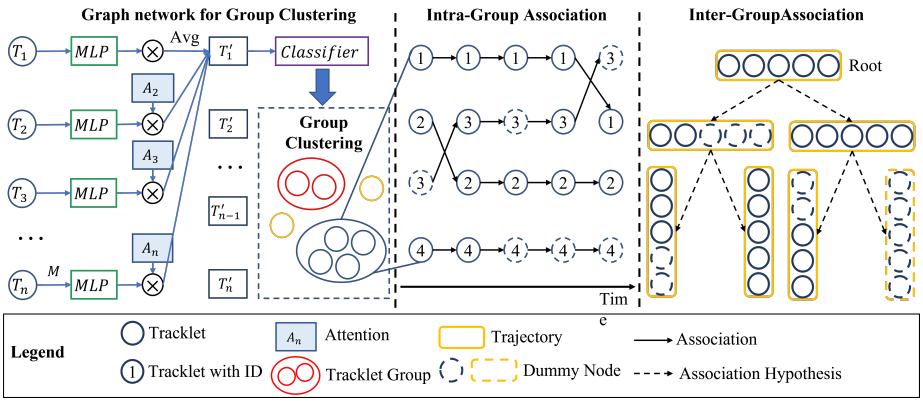


Fig. 2. The general framework of group-guided data association. (1) Initial tracklets are clustered into tracklet groups by the graph network. (2) Intra-group tracklets are associated with the min-cost network flow model. (3) Inter-group trajectories are associated by solving the hypotheses tree.

group is not fully utilized. Chen. et al[6] proposed an online learned elementary grouping model for multi-target tracking. Chen. et al[5] proposed PSTG-based multi-label optimization for multi-target tracking. This method uses the group structure for tracking, but the method is only designed for pairwise target modeling. For tracking problems in more complex scenes, the grouping method of such methods is not robust enough. Sadeghian.[26] using interaction model of object position for affinity measurement. Liu.[18] associates objects by graph matching which considers group structure in measurement. These methods model the relative position structure and do not use the group as the unit for the association. Based on previous research, we use the group structure to recover detection and achieve reliable tracklets, and on the other hand, we use groups as a unit to conduct intra-group and inter-group association of two steps. The method makes full use of group information and achieves accurate tracking in dense scenes.

3 Proposed Method

As shown in Fig. 2, the proposed method consists of three steps. Initial tracklets $\{T_1, T_2, \cdots, T_n\}$ provide motion information for the graph network. Through embedding aggregation and classification, the network gives the grouping score between tracklets. By measuring scores with the threshold, tracklets can be clustered into groups.

To build the optimization model across multiple frames, we use the minimum cost network flow model for the intra-group association. We propose an algorithm to estimate the maximum number of target IDs and only use motion metrics to achieve better computational performance. The dummy nodes in the association graph recover a large number of candidate missed detection.

To obtain long trajectories and eliminate false positives in dummy nodes, we take each trajectory as the root node, establish the association hypothesis and solve the optimal branch. The inter-group association can further reduce the fragments of trajectories and provide long-distance modeling capability for the method.

3.1 Graph Network for Group Clustering

The structure of graph network for group clustering is shown in Fig. 2. To model data association as a graph $G = (V, E)$, vertex set V is consist of all initial tracklets $\{T_1, T_2, \cdots, T_n\}$ in intra-group association window W_{intra}. For tracklet T_j of length p, motion feature M_j is defined as follows:

$$M_j = \{m_1, m_2, \cdots, m_p\} \tag{1}$$

$$m_i = (\frac{x_i/W - \mu_x}{\sigma_x}, \frac{y_i/H - \mu_y}{\sigma_y}, \frac{w_i/W - \mu_w}{\sigma_w} \frac{h_i/H - \mu_h}{\sigma_h}), i \in [1, p], \tag{2}$$

where m_i denotes motion feature of detection D_i in tracklet T_j. (x_i, y_i) denotes detection coordinates, w_i and h_i are width and height of detection. W and H are the width and height of the image, μ and σ represent the mean value and standard deviation. Normalization based on image size makes it easier to improve training efficiency and network performance. To obtain a fixed feature dimension, if the length $p < W_{intra}$, the algorithm performs linear interpolation to insure an equal length of M.

To improve the expression representation capability, the Multilayer Perceptron (MLP) is used to encode features to 512 dimensions, which consist of two Full Connection layers (FC) and Rectified Linear Unit (ReLU). By embedding aggregation, the graph network uses the context information between tracklets to improve the discrimination. For tracklet T_1, embeddings from all tracklets overlapped with it are averaged to update T_1'. To avoid the problem of over smoothing, we calculate the self-attention to discover the importance of other overlapped tracklets. For tracklet pair (T_i, T_j), we follow the paradigm of GAT[34], the importance E_{ij} of tracklet T_j to tracklet T_i is formulated with shared weight $\mathbf{W}$:

$$E_{ij} = A(\mathbf{W}T_i', \mathbf{W}T_j'), \tag{3}$$

where $A()$ is a single-layer feedforward neural network, which maps the high-dimensional features to a number as the attention coefficient. The attention coefficient E_{ij} is nonlinear expressed by LeakyReLU and normalized to α_{ij} by Softmax. To balance the possible deviation of the attention, the multi-head attention is used in the prediction layer. The output is averaged as follows:

$$T_i' = \frac{1}{K} \sum_{k=1}^{K} \sum_{j \in n} \alpha_{ij}^k T_j', \tag{4}$$

where K is number of multi-head attention and n are first-order neighbor vertices of j. The group score $S_{i,j}$ is cosine distance between T_i' and T_j'. The graph

network gives a cluster probability between 0 and 1. Therefore, we set threshold $t_g = 0.7$, when $S_{i,j} > t_g$, tracklets T_i and T_j are clustered into the same group.

Tracklets initialization and training: Initial tracklets provide the basis for group clustering. Based on the method proposed in [27], we extract initial tracklets by affinity measurement of detections. Considering the camera frame rate and pedestrian moving speed, longer tracklets are needed to provide motion features for group clustering. However, the computational performance and accuracy of the baseline method deteriorate with the increase in tracklet length. To handle this problem, frames in an association window are first extracted every 5 frames to generate tracklets. Then, parallel computation is carried out on these 5-frame-long fragments for speed up.

In training, we minimize the cross-entropy loss over all tracklets between the labeled samples and the prediction. The training data are generated from MOT17[20] and MOT20[10] datasets. The positive samples are obtained by measuring the tracklets in the relevant spatio-temporal region. For tracklet pairs, We calculate and accumulate the relative position changes between detections frame by frame. If the cumulative deviation is less than 50% of the object's average displacement, tracklets are labeled into a group. Furthermore, we also manually marked and corrected some positive samples. We randomly shift bounding boxes and delete detections between tracklets to simulate the deviation and missing of the detector. The ratio of positive and negative samples is 1:3. We train for 5000 iterations with a learning rate $5 \cdot 10^{-4}$, weight decay term 10^{-4} and an Adam Optimizer with β_1 and β_2 set to 0.9 and 0.999, respectively. By searching the parameters in the train set, we obtained the optimal parameters. The number of attention head K is set to 4.

3.2 Intra-Group Data Association

In this section, we introduce the intra-group data association. By analyzing object behavior and training data, we propose group constraints that the relative position and number of objects in the same group remain stable for a short time. This property can provide a basis for data association. For example, as shown in Fig. 1 (a), detections are missing in frame 246, and group constraints can be used to construct dummy detection for recovery.

As shown in Fig. 2, the number of object IDs in the group remains static, which is crucial for recovering missed detection. However, there may be two kinds of errors in the initial tracklets: one is that the same object is divided into two different tracklets, and the other is that two different objects are associated into one tracklet. These problems will lead to errors in the number of object IDs in initial tracklets.

In intra-group association, the method first ensures the maximization of recall rate, so it is needed to obtain the maximum possible number N of objects in the window W_{intra} and provide motion features for dummy nodes. First, we use a 2-frame sliding window to calculate the best Perfect matchings of bipartite graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{C})$ and use the Kuhn-Munkres algorithm to assign ID for detections with match $\mathbf{M}$. The Affinity_measure for edges $\mathbf{E}_i$ is based on deepsort[3]. To

Input: $\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_{W_{intra}}$
Output: $N, \boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_{W_{intra}}$

N=0
foreach $\boldsymbol{D}_i, \boldsymbol{D}_{i+1}$, *where i from* 1 *to* $W_{intra} - 1$ **do**
  $\mathbf{V}_i = \{\boldsymbol{D}_i, \boldsymbol{D}_{i+1}\}$
  $\mathbf{E}_i$ =Full connection from $\boldsymbol{D}_i$ to $\boldsymbol{D}_{i+1}$
  $\mathbf{C}_i$ =Affinity_measure($\mathbf{E}_i$)
  $\mathbf{G}_i = (\mathbf{V}_i, \mathbf{E}_i, \mathbf{C}_i)$
  $\mathbf{M}_i$ =Kuhn-Munkres($\mathbf{G}_i$)
  if $|\mathbf{D}_i| > |\mathbf{D}_{i+1}|$ **then**
    Add $(|\boldsymbol{D}_i| - |\boldsymbol{D}_{i+1}|)$ dummy nodes to $\boldsymbol{D}_{i+1}$
  end
  if $|\mathbf{D}_i| < |\mathbf{D}_{i+1}|$ **then**
    Add $(|\boldsymbol{D}_{i+1}| - |\boldsymbol{D}_i|)$ dummy nodes to $\boldsymbol{D}_i$
  end
  N=Max($|\mathbf{M}_i|, N$)
end

return $N, \boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_{W_{intra}}$

Algorithm 1: Algorithm for maximum number N of object IDs

meet the object number constraint, We add the dummy nodes as mismatched detections replica in the subsequent or previous frames. Afterward, N is set to the maximum number of tracklets with different IDs among all frames. The detailed algorithm is shown in Alg. 1.

To obtain higher tracking accuracy, we use the global optimization association. Considering the group constraints, the min-cost network flow model is suitable for modeling the problem. The association edges between the detections are regarded as the path in the network, and the similarity between the detections denotes the cost of the path, solving the min-cost flow of the network between multiple frames can provide the optimal solution. The linear programming problem is as follows:

$$argmin \sum_{i,j \in D} C_i f_i + C_{i,j} f_{i,j} + C_s f_{s,i} + C_t f_{i,t} \tag{5}$$

$$s.t. \quad f \in \{0, 1\} \tag{6}$$

$$f_i = f_s + \sum_{j \in D} f_{j,i} = \sum_{j \in D} f_{i,j} + f_t \tag{7}$$

$$0 \leq \sum_i f_{s,i} \leq 1 \tag{8}$$

$$0 \leq \sum_i f_{t,i} \leq 1 \tag{9}$$

$$\sum_{i \in D} f_i \leq 1, \tag{10}$$

where $f_{i,j}$ denotes flow from detection D_i to D_j and f_i denotes flow from detection D_i to itself. f_s, f_t denotes the virtual source and sink flow with fixed cost C_s and C_t. The data association edge is determined by whether f is activated, i.e. $f \in \{0, 1\}$. Eq .(7) constrains the independence of each trajectory. Eq .(8) and (9) constrain the flow of each target in the network is activated, so the total

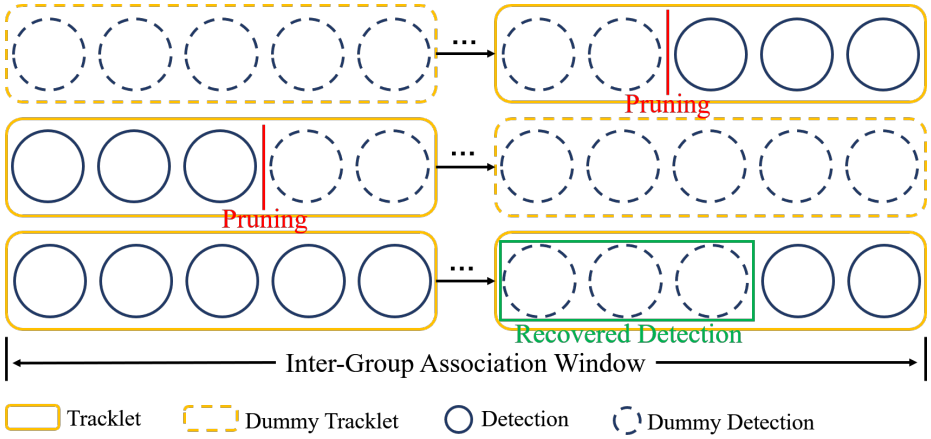


Fig. 3. Pruning and detection recovery example.

flow should be 1. Eq .(10) constrains that two trajectories do not cross one node. Cost C_i denotes the negative confidence of detection D_i, which means the higher the detection confidence, the more likely the detection will be included in the trajectory. Cost $C_{i,j}$ from detection D_i to D_j is as follows:

$$C_{i,j} = GIOU(D_i|D_j), \tag{11}$$

where $GIOU(D_i|D_j)$ denotes generalized intersection over union[24] between detection D_i and D_j. Compared with the traditional Kalman filter, GIOU represents the degree of position similarity and has higher efficiency. Limited by the intra-group association window, the target motion state is relatively stable. Therefore, affinity measurement without appearance feature and linear or nonlinear filter can also obtain high measurement accuracy.

3.3 Inter-Group Data Association

In this section, we introduce the inter-group data association. Due to the group size, the intra-group association cannot model trajectories for a longer time. There are interruptions and fragments caused by window division or occlusion in the trajectories. Trajectories obtained in each group are associated in window W_{inter}. With the increase in the time interval, the error of motion prediction will gradually accumulate. Appearance affinity has become an important basis for tracking associations between groups. Following the tracklet level multiple hypothesis framework[27], the appearance affinity measure $\mathcal{S}$ between tracklets T_i and T_j is as follows:

$$\mathcal{S} = \frac{1}{|T_i|} \sum_{D_i \in T_i} \mathcal{S}_{app}(D_i, D_j), \tag{12}$$

where, D_j denotes the detection in T_j that closest to T_i in time. $\mathcal{S}_{app}$ denotes the cosine distance of appearance feature vectors between detection D_i and D_j. $\mathcal{S}$ averages the appearance of all detections in the T_i which makes it more robust in the face of occlusion. The appearance features of detections are obtained by the Re-Id network [33].

As shown in Fig. 1, there are 2 kinds of nodes in the tracking graph: the original nodes and the dummy nodes. The dummy nodes are the node that is obtained by an extended Kalman filter which predicts the position of the target when it comes to long-term occlusion. Only the original nodes can be used as the root of the hypothesis tree. The association hypothesis increases exponentially with the number of targets, so effective pruning strategies are needed to improve computational efficiency.

Pruning: we take the appearance feature and the space distance as the evaluation branch of the multiple hypothesis tree. However, the correct association may be cut out because of local occlusion, so the algorithm retains the delayed decision to avoid this problem. The dummy detections used by intra-group association can recover the missed detection, but it will cause false-positive problems. We propose a pruning strategy to solve the problem, as shown in Fig. 3. By traversing the branches in the current association hypothesis tree, two kinds of pruning situations are found. One is the branch of continuous dummy detections from the beginning of the window, and the other is the branch of continuous dummy detections from somewhere in the tracklet to the end of the window. We truncate the trajectory according to the position of the red line in Fig. 3. These truncated dummy detections do not match the previous or subsequent objects in the whole calculation window, so they can be regarded as false positives. In addition, as shown in the green box in the figure, the detection of occlusion in the middle is correctly restored, which improves the recall rate of multi-target tracking results.

After pruning, we solve the multi-dimensional assignment problem for the association hypothesis with the strategy proposed in [27] to obtain the final trajectories. The whole MOT method can perform near-online and retain the delay of window size W_{inter}.

4 Experiments

4.1 Dataset and Metric

In experiments, our method is tested on MOT15[17], MOT17[20] and MOT20[10] datasets, which are most widely used in MOT. MOT15 is a comprehensive data set integrating KITTI, ETH, and PETS datasets. MOT17 contains three kinds of public detectors to test the effect of the MOT method on different detectors. Videos are collected from moving and static cameras respectively. MOT20 is designed for ultra-dense scenes, which is more challenging for methods. CLEAR MOT metrics[2] is used to evaluate the method. In addition, IDF1[25] is used to measure the ID accuracy. Higher Order Tracking Accuracy (HOTA)[19] is the

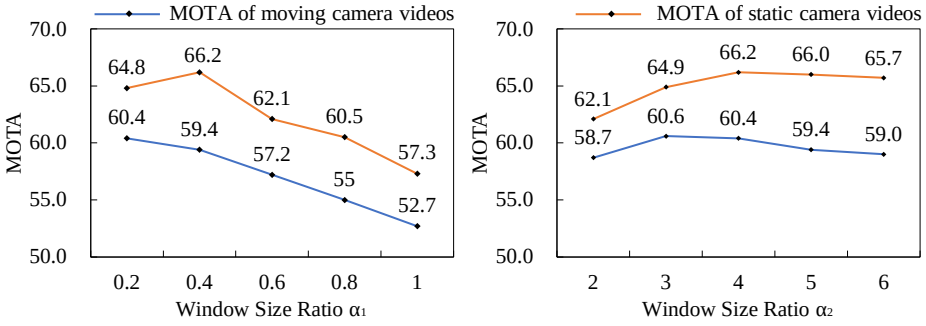


Fig. 4. Window size ratio analysis by moving and static camera videos respectively in MOT17 training set.

Table 1. Comparison of clustering methods on MOT17 validation set

Dataset	Cluster	Setting	mean_Acc ↑	MOTA ↑	IDF1 ↑
MOT17 val	k-means++	k=$\sqrt{No.\ Obj.}$	0.69	34.3	44.7
MOT17 val	agglomerative	Group average	0.84	42.1	53.6
MOT17 val	Our		**0.89**	**43.2**	**55.0**

geometric mean of detection accuracy and association accuracy. Averaged across localization thresholds. For a fair comparison, the tests for all methods use the public detections provided by the dataset.

4.2 Parameters Analysis

The window size of intra-group W_{intra} and inter-group W_{inter} are the main parameters in our method. We searched for the optimal setting of both window sizes on the MOT17 training set. For different window size ratios, Multi-Object Tracking Accuracy (MOTA) of moving and static camera videos in the MOT17 training set are shown in Fig. 4. The window size of the intra-group is related to the object speed and video frame rate, and the window size of the inter-group shall be an integral multiple of the intra-group. Therefore, we define $\alpha_1 = \frac{W_{intra}}{framerate}$ and $\alpha_2 = \frac{W_{inter}}{W_{intra}}$. As shown in Fig. 4, with the increase of α_1, MOTA increased slightly and then decreased because the object structure in the group is no longer stable. Especially in a video with a moving camera, the movement state of the target changes greatly, so it is necessary to use a smaller intra-group window. With the increase of α_2, MOTA increases with the inter-group window, due to the measurement information increases, and the longer trajectory is included in the correlation hypothesis. After α_2 reaches 4, considering the longer time interval of the tracklets is not effective. Therefore, to obtain the best results and achieve a balance between moving and static camera videos, we set $\alpha_1 = 1$ and $\alpha_2 = 4$ for the following experiments.

Table 2. Ablation study on MOT17 training set

Settings	MOTA ↑	IDF1 ↑	FP ↓	FN ↓	IDsw ↓
TT17[41]	56.5	67.0	**9,116**	136,572	572
$A_{intra} + A_{inter}$	58.4	68.7	10,496	131,189	752
$A^{D}_{intra} + A_{inter}$	60.3	70.5	27,245	**100,275**	581
$A^{D}_{intra} + A^{P}_{inter}$	**64.2**	**75.1**	14,232	103,587	**549**

Table 3. Comparison on MOT15 benchmark

Tracker	MOTA ↑	IDF1 ↑	HOTA	FP↓	FN↓	ID Sw. ↓
KCF [7]	38.9	44.5	33.1	7,321	29,501	720
CRFTrack [37]	40.0	49.6	37.3	10,295	25,917	658
Tracktor++v2 [1]	46.6	47.6	37.6	**4,624**	26,896	1,290
Lif_TsimInt [14]	47.2	57.6	43.8	7,635	24,277	554
mfi_tst [39]	49.2	52.4	41.5	8,707	21,594	912
ApLift [15]	51.1	59.0	45.7	10,070	19,288	677
MPNTrack [4]	51.5	58.6	45.0	7,620	21,780	**375**
Our	**57.5**	**60.6**	**46.8**	7,637	**18,013**	466

4.3 Ablation Study

Considering that only group clustering is used without data association, the tracking results will be poor, so we did not use group cluster alone as baseline comparison. First, we made the comparison of cluster methods in the Tab. 1. Heuristic methods often need to manually design and measure feature metrics and search parameters for different scenarios, which is not robust. Graph network has become a mainstream paradigm in the field of deep clustering recently. By embedding aggregation, the context information and high-dimensional features of the graph structure are extracted, which can better model the interaction and motion affinities between trajectories. We separated the validation set from MOT17 and selected common clustering methods for comparison. Without fine-tuning for parameters, we achieve better results on both cluster and tracking.

To verify the effect of each component of the method, we used the MOT17 training set for the ablation experiment. Since our method adopts a multi-stage grouping association strategy based on tracklets, we select a similar tracklet level multi-hypothesis tracking method[41] as the baseline for comparison. As shown in second row of Tab. 2, A_{intra} denotes intra-group association without dummy node, and A_{inter} denotes inter-group association without pruning. We first use A_{intra} and A_{inter} instead of the tracklet generation and association in [41]. Compared with the baseline method in the first row, the basic group association reduces false positives and false negatives and slightly improves main metrics. In the third row of the Tab. 2, A^{D}_{intra} denotes intra-group association with dummy node. Compared with the previous results, the false negative is significantly reduced after the introduction of dummy nodes, indicating that the method recovers a large number of missed detection. However, the dummy

Table 4. Comparison on MOT17 benchmark

Dataset	Method	Detection	MOTA ↑	IDF1 ↑	FP+FN ↓
MOT17 Test	FairMOT[43]	Centernet[11]	73.7	72.3	144,984
MOT17 Test	FairMOT[43]+Group	Centernet[11]	**74.6**	**76.5**	**141,280**
MOT17 Test	ByteTrack[42]	YOLOX[12]	80.3	77.3	109,212
MOT17 Test	Our	YOLOX[12]	**81.0**	**80.0**	**105,668**

Table 5. Comparison on MOT20 benchmark

Method	MOTA ↑	IDF1 ↑	HOTA ↑	FP ↓	FN ↓	IDsw ↓
LPC[9]	56.3	62.5	49.0	**11,726**	213,056	1,562
MPN[4]	57.6	59.1	46.8	16,953	201,384	**126**
ApLift[15]	58.9	56.5	46.6	17,739	192,736	2,241
mfi_tst[39]	59.3	59.1	47.1	36,150	172,782	1,919
TMOH[31]	60.1	61.2	48.9	38,043	165,899	2,342
MPTC[32]	60.6	59.7	48.5	45,318	153,978	4,533
Our	**64.4**	**65.7**	**53.4**	70,976	**110,614**	2,708

introduces wrong estimates resulting in higher false positives. In the fourth row of the Tab. 2, A^P_{inter} denotes inter-group association with pruning. By using the pruning strategy, FP is effectively suppressed and the best result is achieved. The focus of the model affects the preferences of FP and FN, which is a trade-off problem. The bottleneck of the MOT method lies in the recall of the detector, that is FN problem (FN is always one order of magnitude larger than FP in most data sets). Therefore, our strategy is to reduce FN more and keep the sum of FN and FP lower. It is worth mentioning that ID sw. is reduced because pruning reduces unnecessary solution space for the association, thus avoiding ID Sw. caused by similarity measurement ambiguity.

4.4 Benchmark Evaluation

To compare our method with other advanced methods, we chose the classic MOT15, MOT17 benchmark and the most challenging MOT20 benchmark. To fairly compare the performance of data association algorithms, all methods and results use the same detector. In order to prove the generality of the method, we add experiments of group data association method for better detectors and tracking methods. As shown in the Tab. 4, using anchor-free detection and introducing our group association into the popular end to end method FairMOT[43], all the main metrics have been improved. By using same YOLOX[12] detection as ByteTrack, we have achieved SOTA results on MOT17 benchmark. As shown in Tab. 3 and Tab. 5, the best-published results on the leaderboard are listed. Compared with state-of-the-art methods, our method achieves the highest result for MOTA, IDF1, and HOTA. The group-based data association strategy is conducive to achieving higher tracking and identity accuracy. In particular,

Fig. 5. The visual tracking results of MOT20 test set.

our method restores missing detection through group relationships, which has significant advantages in reducing FN. We selected the tracking results of representative frames from the MOT20 test set as a visual display, shown in Fig. 5. It can be observed that in this dense scene, our method can maintain a more stable track ID. The total results can be found on the official website of the MOTChallenge [1].

5 Conclusion

In this paper, we propose a novel group-guided data association for MOT. The graph neural network is designed to obtain the initial groups of tracklets. By analyzing the potential groups of objects, we design a two-stage data association. Intra-group associations utilize the group constraints to achieve more accurate trajectories in dense scenes and recover missed detection. The inter-group association uses the appearance features and proposes tracklets hypothesis to solve the long-term occlusion problem, which improves the trajectory integrity. In the MOT benchmark, the experiments prove the effectiveness of our algorithm, which achieve better results than previous state-of-the-art methods.

Acknowledgements This study is partially supported by the National Key R&D Program of China (No.2019YFB2102200), the National Natural Science

[1] https://motchallenge.net/

Foundation of China (No.61872025), the Science and Technology Development Fund, Macau SAR(File no.0001/2018/AFJ), and the Open Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE2021ZX-03). Thank you for the support from the HAWKEYE Group.

References

1. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 941–951 (2019)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing. pp. 3464–3468. IEEE (2016)
4. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6247–6257 (2020)
5. Chen, J., Sheng, H., Li, C., Xiong, Z.: Pstg-based multi-label optimization for multi-target tracking. Computer Vision and Image Understanding **144**, 217–227 (2016)
6. Chen, X., Qin, Z., An, L., Bhanu, B.: An online learned elementary grouping model for multi-target tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1242–1249 (2014)
7. Chu, P., Fan, H., Tan, C.C., Ling, H.: Online multi-object tracking with instance-aware tracker and dynamic model refreshment. In: IEEE Winter Conference on Applications of Computer Vision. pp. 161–170. IEEE (2019)
8. Chu, P., Ling, H.: Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (October 2019)
9. Dai, P., Weng, R., Choi, W., Zhang, C., He, Z., Ding, W.: Learning a proposal classifier for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2443–2452 (2021)
10. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)
11. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6569–6578 (2019)
12. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
13. Ho, K., Kardoost, A., Pfreundt, F.J., Keuper, J., Keuper, M.: A two-stage minimum cost multicut approach to self-supervised multiple person tracking. In: Proceedings of the Asian Conference on Computer Vision (2020)
14. Hornakova, A., Henschel, R., Rosenhahn, B., Swoboda, P.: Lifted disjoint paths with application in multiple object tracking. In: International Conference on Machine Learning. pp. 4364–4375. PMLR (2020)

15. Hornakova, A., Kaiser, T., Swoboda, P., Rolinek, M., Rosenhahn, B., Henschel, R.: Making higher order mot scalable: An efficient approximate solver for lifted disjoint paths. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6330–6340 (2021)
16. Kratz, L., Nishino, K.: Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. IEEE transactions on pattern analysis and machine intelligence **34**(5), 987–1002 (2011)
17. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)
18. Liu, Q., Chu, Q., Liu, B., Yu, N.: Gsm: Graph similarity model for multi-object tracking. In: IJCAI. pp. 530–536 (2020)
19. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision **129**(2), 548–578 (2021)
20. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
21. Milan, A., Schindler, K., Roth, S.: Multi-target tracking by discrete-continuous energy minimization. IEEE transactions on pattern analysis and machine intelligence **38**(10), 2054–2068 (2015)
22. Mykheievskyi, D., Borysenko, D., Porokhonskyy, V.: Learning local feature descriptors for multiple object tracking. In: Proceedings of the Asian Conference on Computer Vision (2020)
23. Pellegrini, S., Ess, A., Gool, L.V.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: Proceedings of the European conference on computer vision. pp. 452–465. Springer (2010)
24. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
25. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European conference on computer vision. pp. 17–35. Springer (2016)
26. Sadeghian, A., Alahi, A., Savarese, S.: Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 300–311 (2017)
27. Sheng, H., Chen, J., Zhang, Y., Ke, W., Xiong, Z., Yu, J.: Iterative multiple hypothesis tracking with tracklet-level association. IEEE Transactions on Circuits and Systems for Video Technology **29**(12), 3660–3672 (2018)
28. Sheng, H., Lv, K., Liu, Y., Ke, W., Lyu, W., Xiong, Z., Li, W.: Combining pose invariant and discriminative features for vehicle reidentification. IEEE Internet of Things Journal **8**(5), 3189–3200 (2020)
29. Sheng, H., Wang, S., Zhang, Y., Yu, D., Cheng, X., Lyu, W., Xiong, Z.: Near-online tracking with co-occurrence constraints in blockchain-based edge computing. IEEE Internet of Things Journal **8**(4), 2193–2207 (2020)
30. Sheng, H., Zhang, Y., Wang, W., Shan, Z., Fang, Y., Lyu, W., Xiong, Z.: High confident evaluation for smart city services. Frontiers in Environmental Science p. 1103 (2022)
31. Stadler, D., Beyerer, J.: Improving multiple pedestrian tracking by track management and occlusion handling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10958–10967 (2021)

32. Stadler, D., Beyerer, J.: Multi-pedestrian tracking with clusters. In: IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 1–10. IEEE (2021)
33. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision. pp. 480–496 (2018)
34. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
35. Wang, S., Sheng, H., Yang, D., Zhang, Y., Wu, Y., Wang, S.: Extendable multiple nodes recurrent tracking framework with rtu++. IEEE Transactions on Image Processing **31**, 5257–5271 (2022)
36. Wang, S., Sheng, H., Zhang, Y., Wu, Y., Xiong, Z.: A general recurrent tracking framework without real data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13219–13228 (2021)
37. Xiang, J., Xu, G., Ma, C., Hou, J.: End-to-end learning deep crf models for multi-object tracking deep crf models. IEEE Transactions on Circuits and Systems for Video Technology **31**(1), 275–288 (2020)
38. Xu, Y., Chen, Y., Zhang, Y., Zhu, Q., He, Y., Sheng, H.: Bilateral association tracking with parzen window density estimation. IET Image Processing (2022)
39. Yang, J., Ge, H., Yang, J., Tong, Y., Su, S.: Online multi-object tracking using multi-function integration and tracking simulation training. Applied Intelligence pp. 1–21 (2021)
40. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2008)
41. Zhang, Y., Sheng, H., Wu, Y., Wang, S., Lyu, W., Ke, W., Xiong, Z.: Long-term tracking with deep tracklet association. IEEE Transactions on Image Processing **29**, 6694–6706 (2020)
42. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. arXiv preprint arXiv:2110.06864 (2021)
43. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021)
44. Zhao, X., Gong, D., Medioni, G.: Tracking using motion patterns for very crowded scenes. In: Proceedings of the European Conference on Computer Vision. pp. 315–328. Springer (2012)