Iterative Multiple Hypothesis Tracking With Tracklet-Level Association

Hao Sheng, Jiahui Chen, Yang Zhang, Wei Ke, Zhang Xiong, and Jingyi Yu

***Abstract*— This paper proposes a novel iterative maximum weighted independent set (MWIS) algorithm for multiple hypothesis tracking (MHT) in a tracking-by-detection framework. MHT converts the tracking problem into a series of MWIS problems across the tracking time. Previous works solve these NP-hard MWIS problems independently without the use of any prior information from each frame, and they ignore the relevance between adjacent frames. In this paper, we iteratively solve the MWIS problems by using the MWIS solution from the previous frame rather than solving the problem from scratch each time. First, we define five hypothesis categories and a hypothesis transfer model, which explicitly describes the hypothesis relationship between adjacent frames. We also propose a polynomial-time approximation algorithm for the MWIS problem in MHT. In addition to that, we present a confident short tracklet generation method and incorporate tracklet-level association into MHT, which further improves the computational efficiency. Our experiments on both MOT16 and MOT17 benchmarks show that our tracker outperforms all the previously published tracking algorithms on both MOT16 and MOT17 benchmarks. Finally, we demonstrate that the polynomial-time approximate tracker reaches nearly the same tracking performance.**

***Index Terms*— Multiple object tracking, tracking-by-detection, multiple hypothesis tracking, iterative maximum weighted independent set, polynomial-time approximation.**

I. Introduction

MULTIPLE object tracking estimates the spatio-temporal trajectories of targeted, specific objects in video sequences. It is widely used in a variety of applications such as video surveillance [1], human-computer interaction [2], transportation management [3], etc. Although significant progress was made in these areas, there are still existing problems such as heavy occlusions, complex background, and illumination variations that were not completely solved [4], especially in crowd scenes where the targets are frequently either partially or fully occluded. As such, visual tracking still remained a challenge to resolve.

Multiple hypothesis tracking(MHT) is one of the earliest successful tracking methods that was proposed by Reid [5]. Since the main drawback of MHT was the exponential growth of hypotheses, the trackers attempt to maintain tracklet hypotheses effectively through calculating the likelihood of each tracklet and removing hypotheses with low confidence. The best combination of tracklet hypotheses were calculated to estimate the trajectories of multiple objects in each frame, and the best combination finding problem was formulated as Maximum Weighted Independent Set(MWIS) problem. Pruning was usually applied in order to keep the MWIS calculable, and the performance of MHT heavily relies on hypothesis pruning, which needs manual assumptions. The MWIS problem was proven to be NP-hard [6], and the computation time cannot be easily bounded. As such, MHT was considered to be impractical for visual tracking.

With the advances in object detection and feature representation, the tracking-by-detection paradigm was one of the most popular frameworks in multiple object tracking. Target hypotheses were extracted in advanced from video sequences using object detectors, and then object hypotheses were linked by tracking approaches to form trajectories. Multiple object tracking was converted into a data association problem, and the tracker was designed to find an optimal solution to produce final tracking results.

MHT also benefits from both tracking-by-detection paradigm and effective feature representation, and proved to be practical in current visual tracking context in recent years [7]. By incorporating robust convolutional-neural-network-based features and motion information, the number of tracklet hypotheses was significantly reduced which in turn allowed for less assumptions about tracking and explored a larger hypothesis space. MHT converted the tracking problem into a series of MWIS problems across the tracking time, but previous works solved these problems independently.

In this paper, we present a novel iterative MWIS algorithm for MHT. First, we define five tracklet hypothesis categories and the hypothesis category transfer model between adjacent frames. Following this, we propose an iterative algorithm to deal with a series of MWIS problems in MHT that takes

Manuscript received May 22, 2018; revised September 18, 2018 and November 2, 2018; accepted November 3, 2018. Date of publication November 13, 2018; date of current version December 6, 2019. This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1002000, in part by the National Natural Science Foundation of China under Grant 6181101065, in part by the Macao Science and Technology Development Fund under Grant 138/2016/A3, in part by the Open Fund of the State Key Laboratory of Software Development Environment under Grant SKLSDE-2017ZX-09, in part by the Project of Experimental Verification of the Basic Commonness and Key Technical Standards of the Industrial Internet Network Architecture, in part by the Technology Innovation Fund of China Electronic Technology Group Corporation, and in part by the HAWKEYE Group. This paper was recommended by Associate Editor A. Prati. *(Corresponding author: Jiahui Chen.)*

H. Sheng, J. Chen, Y. Zhang, and Z. Xiong are with the School of Computer Science and the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China (e-mail: shenghao@buaa.edu.cn; chenjh@buaa.edu.cn; yang.zhang@buaa.edu.cn; xiongz@buaa.edu.cn).

W. Ke is with School of Public Administration, Macao Polytechnic Institute, Macao 999078, China (e-mail: wke@ipm.edu.mo).

J. Yu is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, and also with the Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA (e-mail: yujingyi@shanghaitech.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2018.2881123

both previous MWIS solution and hypothesis category transfer model into consideration. In addition to that, we also present a polynomial-time approximation algorithm by converting the MWIS problem in a subset of hypotheses into a bipartite graph matching problem. In our formulation, the tracklet hypothesis relevance between adjacent frames is explicitly described and applied to improve the tracking performance and efficiency. Our experimental results on the MOT16 and MOT17 benchmarks demonstrate that our tracker is comparable to published state-of-the-art trackers.

In summary, this paper makes the following contributions:

(1) Presents a new concept, tracklet hypothesis categories, and the category transfer model, which explicitly describes the relevance of tracklet hypotheses between adjacent frames.

(2) Proposes a novel iterative MWIS algorithm, which uses previous MWIS solution and avoids to solve MWIS problem from scratch at each frame with our proposed tracklet category transfer model. It also significantly improves the MWIS solving process in terms of speed and accuracy.

(3) Proposes a polynomial-time approximation algorithm for the MWIS problem in MHT in order to solve the high time complexity problem.

(4) Incorporates a tracklet-level association pruning method into MHT to improve the computational efficiency.

(5) Demonstrates that our tracker outperforms all the previously published tracking algorithms on both MOT16 and MOT17 benchmarks.

II. Related Work

A. Multiple Object Tracking

Tracking-by-detection is a recent standard paradigm for multiple object tracking [8]–[10], which is converted into a data association problem within this framework. The tracking-by-detection approaches can be categorized into recursive and non-recursive methods.

Recursive methods are usually applied to real-time applications because they sequentially build trajectories based on the frame-by-frame associations using the information provided by only previous frames and current frame. Practical Kalman filter based trackers [5], [11] belong to this category. However, these methods tend to produce fragmented trajectories and drift under occlusion and detection errors, because it is more difficult to handle inaccurate detections(e.g. false positives and false negative) compared to the non-recursive methods.

Non-recursive methods [12]–[14] utilize the detections of all sequence frame together to build long tracks robustly against occlusion and inaccurate detections. In general, given a set of detections, short tracklets are generated first by linking individual detections, and then the tracklets are globally associated into long tracklets [15]. As such, global association in these approaches is very important, and many methods for the global association have been proposed [16]–[18] recently. However, the performance of the non-recursive methods is still limited in crowded scenarios. Since these methods usually require the detections for the whole sequence beforehand in addition to heavy computation to generate globally optimized tracks, it is hard to apply the non-recursive methods in real-time applications. Our research belongs to non-recursive methods.

Milan *et al.* [19] introduced a conditional random field to model different types of information jointly for multiple target tracking, including appearance, motion. Butt and Collins [20] proposed a min-cost flow based method to handle the motion model of targets. Chari *et al.* [21] proposed a pairwise cost to enforce or penalize tracklets. Dehghan *et al.* [22] proposed a hierarchical schema to form the tracks iteratively. The cost functions of all of the aforementioned approaches contained only unary and pairwise terms, which were restrictive in modeling high-order information.

MHT permits complex constraints by converting the tracking tasks into multiple dimensional assignments [23], [24]. Since it was too slow and consumed too much memory, MHT was considered impractical in current tracking task. However, with the development of object detection technology and feature representation, MHT has become more practical.

Kim *et al.* [7] incorporated long-term appearance modeling into multiple hypothesis tracking, in which the tracker estimated the online appearance feature for each tracklet. In this method, only detections coupled with the previous tracklet were allowed to updated the tracklets. This reduced the number of hypotheses in order to better simplify manual constraints on tracking so as to make the algorithm more practical. This method focused on filtering tracklet hypotheses through more accuracy features instead of improving MHT itself. Chen *et al.* [25] proposed an enhancing detection model that introduced new conflict constraints to the tracking tasks, by considering full detection information, including the detection-scene model and the detection-detection model. However, additional constraints made the MWIS problem even more complex.

B. Maximum Weighted Independent Set

MWIS had been explored to solving tracking problem within the computer vision community. Papageorgiou and Salpukas [6] converted the tracking tasks into a data association problem and optimized it as an MWIS problem. The method applied an n-scan sliding window to maintain the element number, and then use the maximum weighted independent set algorithm to solve the problem from scratch in each frame.

Brendel *et al.* [26] also introduced MWIS into visual tracking. In this particular work, detection pairs were extracted from two consecutive frames and were used to build the graph. This work focused on model tracking constraints into MWIS model and proved that MWIS was suitable for tracking purpose. However, the special property of MWIS in MHT continues to be ignored.

In mathematics, the MWIS problem was often reformulated as the maximum weight clique(MWC) problem that uses a dual graph of the original [27]. Since the traditional algorithm had a drawback of high space complexity and time complexity, it was not suitable to use on massive graph. Rossi and Ahmed [28] proposed the MCP algorithm for those that relied on the k-Core. Cai [29] proposed a heuristic search method to make the local search even more efficient. Research in the mathematics community focused on solving the massive graph problem from scratch [30], because researches in mathematics field addressed the problem that solving a massive graph only once from scratch.

This paper focuses on solving a series of MWIS problem in MHT. Sec.III proposes a special property of the MWIS problem in tracking problem. Compared to the traditional MWIS problem, the MWIS problems in MHT need to be solved across the the tracking time instead of solving these problems independently from scratch. MWIS problems between two adjacent frames are highly relevant.

III. Iterative Multiple Hypothesis Tracking

This work adopts the tracking-by-detection paradigm [7]. The detections of each frame are given in pre-processing. The detection set is denoted as D. In order to fairly compare against published trackers, we used the public object detection responses on the MOT16 and MOT17 benchmarks [31].

Reid [5] proposed the earliest multiple hypothesis tracking(MHT) framework. Delayed data association decision is the key strategy of MHT, which generates multiple track hypotheses corresponding to one object, and resolves data association ambiguities when further information is obtained. MHT consists of three processes:

(i) **Tracklet Hypothesis Updating:** the tracklet hypothesis updating process maintains multiple possible trajectories for one target. At each frame, new object observations are assigned to existing tracklet hypotheses within tracklet-detection based gating, i.e., motion and appearance.

(ii) **Global Hypothesis Generation:** this resolves the data association ambiguities and finds the best tracklet hypothesis combination, which is explained in a physically plausible way. The selection problem is formulated as maximum weighted independent set (MWIS) problem, known to be NP-hard.

(iii) **Tracklet Hypothesis Pruning:** as the number of tracklet hypotheses exponentially grows, tracklet hypothesis pruning helps to keep the algorithm practical.

MHT algorithm is challenging due to its two opposed objectives: Storing multiple hypotheses for one target until sufficient information is gained to make confident decisions, while simultaneously maintaining as few tracklet hypotheses as possible to keep the method efficient regarding both computation and memory. Both tracklet hypothesis updating and tracklet hypothesis pruning try to strike the right balance between these two constraints. In addition, solving the NP-hard tracklet hypothesis selection problem in global hypothesis tracking in an efficient way is also challenging. In this paper, we address the aforementioned problems.

A. Notation

Before the technical details are provided, we first introduce the notation. $D = \{D^1, \dots, D^t, \dots, D^T\}$ is the set of all input detections, and T is the frame number of the video sequence. All $D(t)$ detections in frame t are represented by the detection set $D^t = \{d_1^t, d_2^t, \dots, d_{D(t)}^t\}$, where d_k^t means the k-th detection at frame t.

TABLE I

Notation. Each Block Summarizes the Symbols for Video Sequence, Detections, Multiple Hypothesis Tracking

Symbol	Description
T	frame number of the video
$\mathcal{D}$	detection set of all frames
D^t	detection set of frame t
$D(t)$	detection number in frame t
d_k^t	the k-th detection in frame t
f_k^t	frame number of detection d_k^t
V	tracklet hypothesis
V_k^t	the k-th tracklet hypothesis in frame t
N^t	number of tracklet hypotheses in frame t
G^t	global hypothesis in frame t
G_k^t	the k-th tracklet hypothesis of global hypothesis in frame t
n^t	number of tracklet hypotheses in global hypothesis in frame t

The detection sequence $(d^1, d^2, \dots, d^k)$ is defined as a **tracklet hypothesis** at frame k, and $d^t \in D^t \cup \{\emptyset\}$. When $d^t \in D^t$, it means detection $d^t \in D^t$ is selected at frame t in this tracklet; when $d^t = \emptyset$, it means that a dummy detection is assigned to this tracklet to deal with a missing detection. Note that for the notational convenience, we make all the tracklet hypotheses from the first frame, but the actual tracklet starts from the first actual detection. Therefore, **global hypothesis** is defined as a set of tracklet hypotheses that are not in conflict; for instance, tracklets should not share detections.

Let $V^t = \{V_1^t, V_2^t, \dots, V_{N^t}^t\}$ be the tracklet hypothesis set at frame t, and N^t is the hypothesis number. Then $G^t = \{G_1^t, G_2^t, \dots, G_{n^t}^t\}$ is the global hypothesis at frame k, whose elements are selected in the current global hypothesis.

B. Framework Overview

The tracker maintains tracklet hypotheses for tracked targets and calculates the likelihood of each tracklet. Following this, the best combination of tracklet hypotheses is found to estimate the trajectories of multiple objects. However, the number of potential hypotheses grows exponentially across the tracking time, and the performance of MHT heavily relies on hypothesis pruning, which needs manual assumptions. The combination selection problem is formulated as the maximum weighted independent set(MWIS) problem, and that is proven to be NP-hard [6]. As such, the computation time cannot be easily bounded,so the MHT is considered to be impractical for visual tracking. This section introduces the traditional multiple hypothesis tracking under the framework of tracking-by-detection.

MHT constructs tracklet hypotheses for all potential trajectories, and updates these hypotheses in a frame-by-frame manner. MHT attempts to find a most-likely global hypothesis from all tracklet hypotheses at each frame for either pruning purpose or the trajectory generation. It consists of three processes:

1) Tracklet Hypothesis Updating: MHT maintains possible tracklets from the first frame to the current frame. At each frame, tracklet hypotheses are constructed to represent new tracklets starting from current detections. Existing tracklet hypotheses also need to be updated. (1) Previous tracklet hypotheses are extended by dummy detections to account for missing detections, and (2) the previous tracklet hypotheses

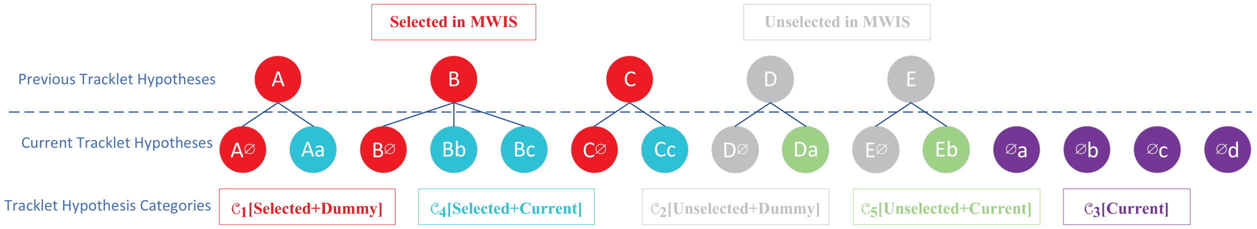

Selected in MWIS
Unselected in MWIS
Previous Tracklet Hypotheses
A
B
C
D
E
Current Tracklet Hypotheses
A∅
Aa
B∅
Bb
Bc
C∅
Cc
D∅
Da
E∅
Eb
∅a
∅b
∅c
∅d
Tracklet Hypothesis Categories
$\mathcal{C}_1$[Selected+Dummy]
$\mathcal{C}_4$[Selected+Current]
$\mathcal{C}_2$[Unselected+Dummy]
$\mathcal{C}_5$[Unselected+Current]
$\mathcal{C}_3$[Current]

Fig. 1. Hypothesis Category Transfer. Uppercase letters represent tracklets; lowercase letters represent detections; Ø represents dummy detections. The combination of a uppercase letter and a lowercase letter(or Ø) means a tracklet updated with a detection(or dummy detection). Note that the combination of Ø and a lowercase letter means a new tracklet starting with a detection. Current tracklet hypotheses are divided into five categories according to Table II.

are also updated with current detections, which are similar to the tracklets in some aspects, i.e., appearance and motion.

2) Global Hypothesis Generation: Throughout this process, a subset of tracklet hypotheses is selected to form a global hypothesis. Before the selection, the score of tracklet hypotheses needs to be calculated, and then the best selection problem is solved as an MWIS problem.

We construct an undirected graph $G = (V, E)$ to model the MWIS. Each vertex denotes one tracklet hypothesis h with weight w_h. When the two tracklet hypotheses i and j are in conflict, an edge $(i, j) \in E$ connects these two vertices. Then MWIS problem is formulated as follows:

$$\max_x \ \Sigma_h w_h x_h \tag{1}$$

$$s.t. \ x_i + x_j <= 1, \quad (i, j) \in E \tag{2}$$

$$x_i \in \{0, 1\}, \tag{3}$$

where x_h is an indicator. When $x_h = 1$, it means that hypothesis h is selected in the current global hypothesis; when $x_h = 0$, it means that the hypothesis h is not selected. Eq. 3 represents that two tracklets cannot be simultaneously selected when they are in conflict.

3) Tracklet Hypothesis Pruning: Since the number of tracklet hypotheses grows exponentially, tracklet hypothesis pruning is an essential process that reduces the tracklet hypotheses which deviate significantly from the current global hypothesis.

To summarize, traditional MHT updates tracklet hypotheses at each frame, and seeks the best feasible combination known as global hypothesis independently. Note that MWIS problem is an NP-hard problem, which is both time-consuming or sub-optimal and leads to inaccurate tracking result.

Given that tracklet hypothesis set V^k and global hypothesis G^k at frame k, the goal is to generate a new tracklet hypothesis set V^{k+1} at frame $k+1$ and to effectively solve the global hypothesis generation problem to find the global hypothesis G^{k+1}. Note that for the first frame, V^1 is the tracklet hypothesis set, whose elements are tracklets with only one detection of D^1. G^1 selects all the elements in V^1 to form the global hypothesis, since no detections are shared by tracklets.

C. Tracklet Hypothesis Updating and Tracklet Category Transfer Model

V^k and G^k are respectively the tracklet hypothesis set and the global hypothesis set at frame k. The elements of V^k can be divided into two categories as shown in Fig. 1. The red nodes in previous tracklet hypotheses represent the selected tracklet hypotheses which are elements of last MWIS solution G^k, as well as the gray nodes in previous tracklet hypotheses represent the unselected tracklet hypotheses which are elements of $V^k \setminus G^k$. Given detection set D^{k+1} at frame $k+1$, then the tracklet hypotheses are updated as follows:

(1) Previous tracklet hypotheses are extended by appending dummy detections to represent the tracklets with missing detections at frame $k+1$. The dummy detections are used to account for the missing detections. We define that the selected hypotheses(G^k) updated with dummy detections belong to the category $\mathcal{C}_1$, and the unselected hypotheses($V^k \setminus G^k$) updated with dummy detections belong to the category $\mathcal{C}_2$. In Fig. 1, hypotheses of $\mathcal{C}_1$ are in red, and hypotheses of $\mathcal{C}_2$ are in gray.

(2) For each detection in D^{k+1}, we create a new tracklet hypothesis that represents a new target entering the scene and we define that these hypotheses are in the category of $\mathcal{C}_3$. In Fig. 1, hypotheses of $\mathcal{C}_3$ are in purple.

(3) Previous tracklet hypotheses are updated with current detection set D^{k+1}. Note that each tracklet hypothesis only updates with detections that are reasonable regarding motion and appearance. We define that the selected hypotheses(G^k) updated with current detections belong to the category $\mathcal{C}_4$, and the unselected hypotheses($V^k \setminus G^k$) updated with dummy detections belong to the category $\mathcal{C}_5$. In Fig. 1, hypotheses of $\mathcal{C}_4$ are in blue, and hypotheses in $\mathcal{C}_5$ are in green.

Given tracklet hypothesis set V^k and global hypothesis set G^k at frame k, tracklet hypothesis set V^{k+1} is updated with current detection set D^{k+1} or dummy detections. Based on the original tracklet hypothesis and current detection, we divide new tracklet hypotheses into five categories as shown in Table II.

D. Tracklet Hypothesis Reduction

Tracklet hypothesis reduction is used to reduce the number of tracklet hypotheses when a feasible solution is found. We decrease the number of tracklet hypotheses and make the MWIS problem more likely to get an exact solution instead of an approximation solution. Tracklet hypothesis reduction also speeds up the process.

We first define two notations, $F(v)$ and $F[v]$. The feasible tracklet hypothesis set of tracklet v at frame k is defined as follows:

$$F(v) = \{u \in V^k | u \text{ and } v \text{ is not in conflict}\}, \tag{4}$$

TABLE II

TRACKLET HYPOTHESIS CATEGORY

Symbol	Description
$\mathcal{C}_1$	**[Selected+Dummy]** The tracklet hypothesis selected in global hypothesis G^k is updated with dummy detections.
$\mathcal{C}_2$	**[Unselected+Dummy]** The tracklet hypothesis unselected in global hypothesis G^k is updated with dummy detections.
$\mathcal{C}_3$	**[Current]** Current detection is regarded as start of a new tracklet.
$\mathcal{C}_4$	**[Selected+Current]** The tracklet hypothesis selected in global hypothesis G^k is updated with current detections.
$\mathcal{C}_5$	**[Unselected+Current]** The tracklet hypothesis unselected in global hypothesis G^k is updated with current detections.

where V^k is the tracklet hypothesis set at frame k. Only elements in $F(v)$ can be selected in global tracklet hypothesis at frame k, when tracklet v is selected. Moreover, we denote $F[v] = F(v) \cup \{v\}$.

In order to reduce the number of hypotheses, we estimate the weight upper bound for each hypothesis and remove the hypotheses whose upper bound is less than the weight of the current solution. Let $UB(v)$ be the upper bound of the weight of independent set containing v.

Let n^* be the tracklet hypothesis with the maximum weight in $F(v)$, then $UB(v)$ is calculated as follows:

$$UB(v) = \max\{w(F[v]) - w(n^*),\\ w(v) + w(n^*) + w(F(v) \cap F(n^*)\}, \quad (5)$$

where w is the weight function, which is the sum of weights of all the elements. We calculate the upper bounds for two cases: When n^* is not selected, the feasible tracklet hypothesis candidate set is $F(v) = F([v]) \setminus \{n^*\}$, and the upper bound is $w(F[v]) - w(n^*)$; when n^* is selected, the feasible tracklet hypothesis candidates should be $F(v) \cap F(n^*)$ and the upper bound should be $w(v) + w(n^*) + w(F(v) \cap F(n^*))$.

$UB(v)$ is the upper bound of weight when tracklet hypothesis v is selected in the solution. So if a tracklet hypothesis v with $UB(v)$ less than the weight current feasible solution, there is no chance the tracklet hypothesis v is selected in the solution. So we remove the hypothesis v from our hypothesis set and $UB(v)$ is the evidence of our hypothesis reduction before MWIS is applied.

Once we get a feasible solution, we apply our tracklet hypothesis reduction to simplify the MWIS problem, as shown in Alg.1. First, we calculate the upper bounds for each hypothesis(line 2-5). Following this, we remove the unlikely hypotheses from our tracklet hypothesis set(line 9). Note that once a tracklet hypothesis u is removed, we update the upper bound of $F(u)$ and find new unlikely hypotheses for further removing(line 10-14).

IV. TRACKLET-LEVEL ASSOCIATION

MHT tracks multiple objects in a frame-by-frame manner. In each frame, the current detections are assigned to existing tracklets or regarded as newly initiated tracklet. In order to maintain efficiency, a tracklet-detection gating is usually applied when assigning detections to existing tracklets, so that only the tracklet-detection association is considered in this process. The detections produced by object detector are noisy, so the MHT has to store a large number of tracklet hypotheses to make the actual track. The mechanism leads to the efficiency problem. In this work, we incorporate the tracklet-level association in the tracklet hypothesis updating process.

Algorithm 1 Tracklet Hypothesis Reduction

```
Input: Tracklet hypothesis set $V$, a feasible solution $S_0$
Output: Reduced tracklet hypothesis set $V$
 1: for each $v \in V$ do
 2:   if $UB(v) \le w(S_0)$ then
 3:     $RmTracklets \leftarrow RmTracklets \cup \{v\}$
 4:   end if
 5: end for
 6: while $RmTracklets$ is not empty do
 7:   $u \leftarrow$ pop a tracklet from $RmTracklets$
 8:   Remove $u$ from $V$
 9:   for each $v \in V$ do
10:     if $UB(v) \le w(S_0)$ then
11:       $RmTracklets \leftarrow RmTracklets \cup \{v\}$
12:     end if
13:   end for
14: end while
15: return $V$
```

A. Short Tracklet Generation

In frame-by-frame tracking method, only relationships between detections in consecutive frames are taken into consideration when forming the tracklets. The appearance difference gradually increases along with the time gap, to eventually result in an ID switch problem. Moreover, these errors are propagated to the final solution because the ID switch caused by this reason is difficult to aware.

In order to handle this issue, we introduce a net structure for detection association in a temporal domain inspired by [22]. Instead of solely considering the correlation between two consecutive frames, we form the tracklets based on all correlation between any two detections. In order to make this tracklet generation algorithm effective, we split the video sequence into small windows so as to generate the short tracklet within five frames.

B. K-Partite Detection Graph

Given the detection set of a K frame window, without loss of generality, we denote the detection set $D = \{D^1, D^2, \ldots, D^K\}$. Then, we build a k-partite graph $G = (V_1, V_2, \ldots, V_K; E; W)$ to formulate our short tracklet generation as shown in Fig. 2. Relationships between detections in a temporal windows are modeled in the k-partite graph. Note that although all relationships between detections are considered, we only link two detections with high similarity in different frames. In this way, a clique in the k-partite graph

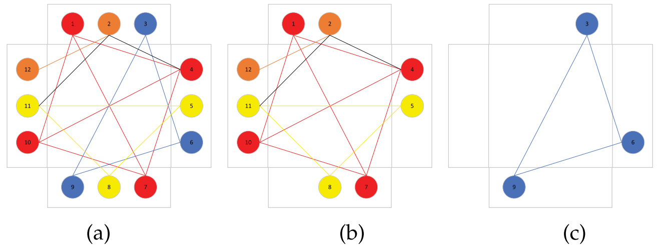


Fig. 2. K-partite Graph for Short Tracklet Generation. (a) is the original k-partite detection graph. Based on connectivity, (a) is divided into two subgraph (b) and (c).

form the track of an object: In a clique, all nodes connect to each other, and any two detections in this clique have high similarity. In order to find the global optimization solution, we simultaneously consider all cliques.

Node Set: For each node set V_i, let $V_i = D^i$. Each node set corresponds to detections in a frame and each node represents a detection.

Edge Set: For each pair of nodes, there is an edge when two detections are similar with each other. Note that for each pair of nodes in one node set, there are no edges.

Weight Set: For each edge, we assign a weight to describe the similarity between two linked detections. Our experiments use the cosine distance of the convolutional-neural-network-based features.

Then we estimate the best cliques, which is formulated as follows:

$$\max_{e_{ij}} \ \Sigma_{e_{ij}} w_{ij} e_{ij} \tag{6}$$

$$s.t. \ e_{ij} \text{ form cliques} \tag{7}$$

$$e_{ij} = 0, \quad \text{sim}(i, j) < \alpha \tag{8}$$

$$e_{ij} \in \{0, 1\} \tag{9}$$

$$e_{ij} = e_{ji} \tag{10}$$

Eq. 6 tells that we use the sum of all edge weights in a clique to evaluate the clique.

Eq. 7 ensures the edges form cliques. We make the constraints as follows: For each three nodes in three different node set should satisfy

$$e_{ij} + e_{jk} <= e_{ik} + 1 \tag{11}$$

The constraint means that for any three nodes i, j, k, if i and j connect to each other and j and k connect to each other, then i and k also needs to be connected.

The similarity used in Eq. 8 is the cosine distance between features of two detections. The detection features is a generated by VGG16 network [32].

Due to the unimodular property, the solution of the optimization problem is all integer solution. We apply linear program method to get the optimization solution.

After we get the solution of the problem, we get e_{ij}. And e_{ij} form a clique set $C = \{C_1, C_2, \ldots, C_n\}$. For each clique C_i in C, C_i contains several nodes in k-partite graph and for each V_i, at most one node is included. In the end, the maximum element number of C_i is k, and all these k nodes form the track of an object and in each frame, the object is extracted by object detector. For other cliques whose element number is less than k, there are two situations: 1) the actual trajectory of the object is less than k frame long, and 2) the object is missed by detector in some frames.

C. Speed-Up

We need to enumerate all three nodes in different clusters due to the clique constraint set, but this leads to a large number of constraints, and low effectiveness. Based on the connectivity of the k-partite graph, we divide the graph into several subgraphs and there is no edges between any subgraphs. As shown in Fig. 2, the original four-partite graph is divided into two subgraphs.

We know that the combination of solution in each subgraph is the solution of the whole k-partite graph. The number of constraints decrease significantly based on Eq. 7.

D. Tracklet Updating With Tracklet-Level Association

The tracklet updating process generates tracklet hypotheses with previous existing tracklet hypotheses and current detections. In a traditional method, detections have to satisfied the gating with a existing tracklet in terms of some aspects, i.e. motion and appearance. Then, the existing tracklet hypotheses is extended by the detection to form new tracklet hypothesis. In this way, only relationship between a tracklet and a detection is considered in this process. Besides traditional distance and motion gating, we also introduce our novel tracklet-level association. As introduced in Sec.IV-A, we first generate the short tracklet set in every k frame window, and these frame windows are non-overlapped, e.g., frame $1 \sim k$, frame $k+1 \sim 2k$. Then we have the short tracklet sets from each window, and each tracklet is regarded as the tracking unit rather than detections as the tracking unit in traditional multiple hypothesis tracking. Then the average position and the average appearance feature vector of all detections are used to describe the motion and appearance information. The tracklet-level association significantly reduces the hypothesis number.

V. Iterative Maximum Weighted Independent Set

Our proposed iterative MWIS algorithm takes advantages of previous MWIS solution instead of solving the problem from scratch each time. Based on our tracklet hypothesis category transfer model, we estimate a feasible solution and simplify current MWIS problem according to tracklet hypothesis reduction introduced in Sec.III-D. The reduction process is applied once we get a better solution.

Alg.2 presents the pseudo code of our method. Aside from current hypothesis candidates, a feasible solution is also required as the input of our algorithm, so we introduce the initial feasible solution generation.

Given the current tracklet hypothesis set V, and we divide V into five categories according to Table I, then $V = \mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3 + \mathcal{C}_4 + \mathcal{C}_5$. One feasible MWIS solution is $\mathcal{C}_1 + \mathcal{C}_3$. As the elements in $\mathcal{C}_1$ correspond to the previous solution, elements in

Algorithm 2 Iterative Maximum Weighted Independent Set

```
Input: Tracklet hypothesis set $V$, a feasible solution $S_0$
Output: Maximum weighted solution $C^*$
1:  $V$ = TH_Reduction($V, S_0$)
2:  while $V$ is not empty do
3:    $u \leftarrow$ tracklet with maximum weight in $V$
4:    $C \leftarrow \{u\}$
5:    $Candidates \leftarrow F(u)$
6:    while $Candidates$ is not empty do
7:      $v \leftarrow$ tracklet with maximum $benefit(v)$
8:      if $w(C) + w(v) + w(F(v) \cap Candidates) \leq w(S_0)$ then
9:        Break
10:     end if
11:     $C \leftarrow C \cup v$
12:     $Candidates \leftarrow Candidates \backslash \{v\}$
13:     $Candidates \leftarrow Candidates \cap F(v)$
14:   end while
15:   if $w(C) > w(C^*)$ then
16:     $C^* \leftarrow C$
17:     $V \leftarrow TH\_Reduction(V, C^*)$
18:     if $V$ is empty then
19:       return $C^*$
20:     end if
21:   end if
22: end while
23: return $C^*$
```

this category do not share any detections. In addition, elements in $\mathcal{C}_3$ correspond to different detections in the current frame, so they are also feasible. Lastly, $\mathcal{C}_1$ and $\mathcal{C}_3$ do not have any detections in common, as the detections in $\mathcal{C}_1$ are in previous frames, while detections in $\mathcal{C}_3$ are current detections. Note that in Sec.VI, we introduce another initial solution generation method.

At the beginning of the algorithm, we remove some hypotheses by applying our tracklet hypothesis reduction with a feasible solution $S_0 = \mathcal{C}_1 + \mathcal{C}_3$. V remains the rest hypothesis candidates(lines 2). Then, the algorithm executes the main loop until an exact solution is found or there is no feasible candidate. In each iteration, we form a feasible combination of tracklet hypothesis(lines 4-16). Once a better solution is found, we update the current solution and remove candidates according to Alg.1.

The score of tracklet hypothesis $\mathcal{T} = (i^1, i^2, \ldots, i^k)$ is defined as follows:

$$\mathrm{sc}(\mathcal{T}) = \sum_{m=1}^{k} \mathrm{Con}(i^m) + \sum_{m=1}^{k-1} \mathrm{Aff}(i^m, i^{m+1}), \tag{12}$$

where $\mathrm{Con}(i^m)$ is the confident of short tracklet and $\mathrm{Aff}(i^m, i^{m+1})$ is the affinity between two tracklets. Eq. 12 is used to calculate the weights of tracklets.

The tracklet-level affinity is defined as follows:

$$\mathrm{Aff}(i^m, i^{m+1}) = \cos\left(\frac{\sum_{d \in i^m} \mathrm{App}(d)}{|i^m|}, \frac{\sum_{d \in i^{m+1}} \mathrm{App}(d)}{|i^{m+1}|}\right), \tag{13}$$

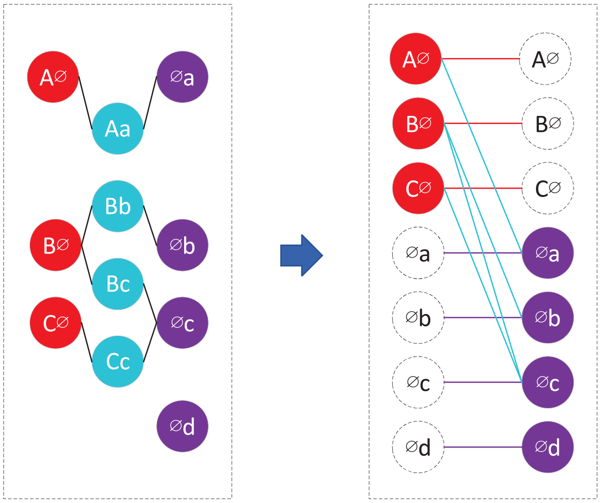


Fig. 3. MWIS Reformulation. Uppercase letters represent tracklets; lowercase letters represent detections; ∅ represents dummy detections. The combination of a uppercase letter and a lowercase letter(or ∅) means a tracklet updated with a detection(or dummy detection). Note that the combination of ∅ and a lowercase letter means a new tracklet starting with a detection.

where $|i|$ is the detection number in tracklet i. The detection features is a generated by VGG16 network [32].

VI. Polynomial-Time Approximation

In the previous section, we present the iterative MWIS algorithm, which simplifies the problem by taking advantage of previous solution and category transfer model. However, the algorithm is still an NP-hard problem and no existing polynomial-time method produces the optimal solution. In this section, we propose a polynomial-time approximation method to solve the MWIS problem in MHT.

We first propose two proposition as follows:

Proposition 1: Given the current tracklet hypothesis set V, and we divide V into five categories according to Table I, then $V = \mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3 + \mathcal{C}_4 + \mathcal{C}_5$. $V_{134} = \mathcal{C}_1 + \mathcal{C}_3 + \mathcal{C}_4$ is the subset of V. Then the maximum weighted independent set problem of V_{134} can be reformulated as the maximum bipartite matching problem.

Proof: Since $\mathcal{C}_1$ is the solution of previous MWIS problem, its elements are independent of each other: any two tracklets do not share detections. In addition, $\mathcal{C}_3$ only contains tracklets with one current detection, so the elements in $\mathcal{C}_1$ and $\mathcal{C}_3$ are independent of each other. $\mathcal{C}_1 + \mathcal{C}_3$ is a feasible solution. Tracklet hypothesis v in $\mathcal{C}_4$ are only in conflict with an element of $\mathcal{C}_1$ and an element of $\mathcal{C}_3$. Assuming there are respectively m, n, k elements in $\mathcal{C}_1, \mathcal{C}_3$ and $\mathcal{C}_4$. ■

Afterwards, we construct a bipartite graph $G = (S, T; E)$ as shown in Fig. 3, where S has m real vertices corresponding to elements in $\mathcal{C}_1$ and n virtual vertices corresponding to elements in $\mathcal{C}_3$; T has m virtual vertices corresponding to elements in $\mathcal{C}_1$ and n real vertices corresponding to elements in $\mathcal{C}_3$; E is defined as follows:

(1) For each tracklet hypothesis h in $\mathcal{C}_1$, we link real vertex corresponding to h in S and virtual vertex corresponding to h with weight $w(h)$. These edges are in red in Fig. 3.

(2) For each tracklet hypothesis h in $\mathcal{C}_3$, we link virtual vertex corresponding to h in S and real vertex corresponding to h with weight $w(h)$. These edges are in purple in Fig. 3.

(3) For each tracklet hypothesis h in $\mathcal{C}_4$, h is corresponding to a tracklet hypothesis h_1 in $\mathcal{C}_1$ and a tracklet hypothesis h_2 in h_3. We link real vertex corresponding to h_1 in S and real vertex corresponding to h_2 with weight $w(h)$. These edges are in blue in Fig. 3.

This way, the MWIS problem is reformulated as a maximum bipartite matching problem, which can be solved by using Hungarian Algorithm in polynomial time.

Proposition 2: Given the current tracklet hypothesis set V, and we divide V into five categories according to Table I, then $V = \mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3 + \mathcal{C}_4 + \mathcal{C}_5$. $V_{1234} = \mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3 + \mathcal{C}_4$ is the subset of V. Then the elements of $\mathcal{C}_2$ are not selected in the maximum weighted independent set.

Proof: As $\mathcal{C}_1 \cup \mathcal{C}_2$ equals to the MWIS problem in last frame, then we have

$$F(\mathcal{C}_1 + \mathcal{C}_2) = w(\mathcal{C}_1), \tag{14}$$

where F is the maximum weight of MWIS in $\mathcal{C}_{12}$ and w is the weight function, which is the sum of weights of all the elements. And according to **Proposition 1**, we have $F(\mathcal{C}_1 + \mathcal{C}_3 + \mathcal{C}_4) = w(\mathcal{C}_1) + w(\mathcal{C}_3) + f(\mathcal{C}_1, \mathcal{C}_3)$, where $f(\mathcal{C}_1, \mathcal{C}_3) \geq 0$ is the weight gain from V_4. Based on Hungarian Algorithm, we have

$$\mathcal{C}_1' \subset \mathcal{C}_1, \mathcal{C}_3' \subset \mathcal{C}_3 \Rightarrow f(\mathcal{C}_1', \mathcal{C}_3') \leq f(\mathcal{C}_1, \mathcal{C}_3) \tag{15}$$

■

Based on the selection of $\mathcal{C}_2$, the maximum weight of V_{1234} can be defined as

$$F(V_{1234}) = \max_{\mathcal{C}_2'} w(\mathcal{C}_1') + w(\mathcal{C}_2') + w(\mathcal{C}_3) + f(\mathcal{C}_1', \mathcal{C}_3), \tag{16}$$

where the selected set $\mathcal{C}_2'$ in $\mathcal{C}_2$ only affect selection in $\mathcal{C}_1$. We know that when $\mathcal{C}_2' = \emptyset$, both terms $w(\mathcal{C}_1') + w(\mathcal{C}_2')$ and $f(\mathcal{C}_1', \mathcal{C}_3)$ are at their maximum according to Eq. 14 and Eq. 15. As such, we prove that elements in $\mathcal{C}_2$ are not be selected.

Based on the two above mentioned propositions, the conclusion is that if we ignore $\mathcal{C}_5$, MWIS problem of $\mathcal{C}_{1234}$ can be solved in polynomial time. Moreover, elements in $\mathcal{C}_2$ are not selected in the current global hypothesis.

As the updated hypotheses from the unselected hypotheses with dummy detections are unlikely to be chosen at the current frame, we assume that the hypotheses updated from the unselected hypotheses with current detections are also unlikely to be chosen at the current frame. The experiments analyze the reasonableness of the assumption. Based on the two aforementioned propositions, we conclude that if we can ignore $\mathcal{C}_5$, MWIS of V_{1234} can be solved in polynomial time. Then our algorithm is given in Alg.3. For each frame, we first update previous selected tracklets update with dummy detections and current detections in order to get new hypothesis set $\mathcal{C}_1$(line 1) and $\mathcal{C}_4$(line 3). New tracklet hypotheses starting from current detections are also generated(line 2). Lastly, we reformulated the WMIS problem as maximum bipartite graph matching problem according to **Proposition 1**. The Algorithm also can be applied to generate initial feasible solution for our iterative MWIS algorithm introduced in Sec.V.

Algorithm 3 Approximation Method

```
Input: Previous solution $S_{pre}, currentdetectionset D_k$
Output: Solution $S$
1: $V_1 \leftarrow S_{pre}$ updating with dummy detections
2: $V_3 \leftarrow$ new tracklets starting from $D_k$
3: $V_4 \leftarrow S_{pre}$ updating with $D_k$
4: $S \leftarrow MWIS(V_1, V_3, V_4)$ according to Proposition 1
5: return $S$
```

VII. Experiments

Dataset: We tested our approach on the MOT16 and MOT17 benchmark [31] and achieved very competitive results. There were seven training sequences and seven test sequences in the MOT16 benchmark, along with twenty-one training sequences and twenty-one test sequences in the MOT17 benchmark. Sec.VII-A gives the details about short tracklet generation. Sec.VII-B demonstrates the evaluation results on the training sequences in order to verify the effectiveness of tracklet hypothesis reduction; Sec.VII-C verified our proposed iterative MWIS algorithm; Sec.VII-D analyzes our tracking performance of the approximation algorithm; Sec.VII-E compares our method with other state-of-the-art tracking methods. In order to maintain a fair comparison, we used the public detection set given by MOT16 and MOT17 as our algorithm input. All tracking approaches are based on the same input.

Evaluation Metric: We follow the standard CLEAR MOT metrics [38] for evaluating tracking performance. The metrics includes multiple object tracking accuracy (MOTA↑), which combines identity switches (ID Sw.↓), false positives (FP↓), and false negatives (FN↓). Beside we also report mostly tracked (MT↑), mostly lost (ML↓) and fragmentation (Frag↓). Researchers usually use multiple object tracking accuracy (MOTA) to compare different trackers. However, it has been pointed out that MOTA does not properly account for identity switches [39], [40], as shown in the left of Fig. 4. More adapted metrics have therefore been proposed. For example, IDF1 is computed by matching trajectories to ground-truth so as to minimize the sum of discrepancies between corresponding pairs [41]. Unlike MOTA, it penalizes switches over the whole trajectory fragments assigned to the wrong identity, as shown in the right side of Fig. 4. In this section, we report results both in terms of MOTA and IDF1, to highlight the drop in identity switches. ↑ is a positive indicator meaning the higher the value, the better, while ↓ means the lower the value, the better.

A. Short Tracklet Verification

In this subsection, we discuss the short tracklet generation. We first give the runtime of the short tracklet generation with different window size, as shown in the Fig. 6(a). All these results are calculated in the most crowded sequence in MOT benchmark, as shown in Fig. 6(b), the process speed is much faster in easy sequences. In this experiment, we change the

TABLE III
SHORT TRACKLET ACCURACY

MOT16			MOT17		
#tracklet	without ID Sw.	with IDS	#tracklet	without ID Sw.	with ID Sw.
10,705	10,514 (98.2%)	191 (1.8%)	37,700	37,290 (98.9%)	410 (1.1%)

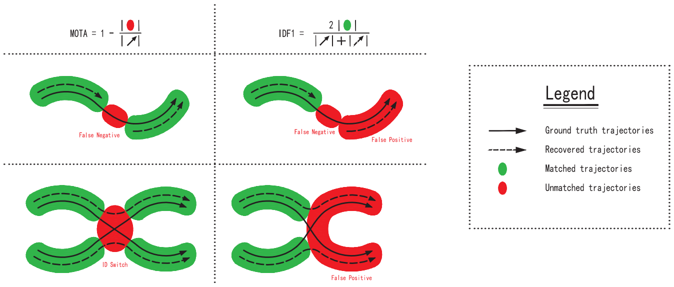


Fig. 4. Effect of ID Switch on tracking overall metrics, MOTA and IDF1. The solid lines represent the ground-truth trajectories and the dotted lines is the recovered trajectories.

TABLE IV
NUMBER OF DIFFERENT HYPOTHESIS REMOVING PERCENTAGE

Dataset	Percentage of Removing Hypotheses				
	100%	75-99%	50-74%	25-49%	0-24%
MOT16	126944 (98.57%)	876 (0.68%)	376 (0.29%)	281 (0.22%)	305 (0.24%)
MOT17	336324 (98.30%)	2367 (0.69%)	1017 (0.30%)	915 (0.27%)	1510 (0.44%)

window size from 3 to 7 and record the runtime of tracklet generation. As the number of constraints grows exponentially with the window size according to Eq. 3, the average process time also increase. To balance the tracklet length and computational efficiency, we set the frame size to 5 in our tracking experiments.

Then we demonstrate the accuracy of short tracklet as shown in Table III. ID switch is the most serious problem in this step because the ID switch error in short tracklet propagates to final trajectories. So we give the total number of short tracklets, as well as, the number of tracklet with and without ID switch in the table. We generated respectively 10,705 and 37,700 short tracklets in MOT16 and MOT17 benchmarks, and more than 98.2% of these tracklets are correct without ID switch errors. It proves that our proposed maximum-multi-clique based method is able to procedure confident tracklets.

B. Hypothesis Reduction

Our first experiment focused on analyzing the effectiveness of hypothesis reduction in our iterative MWIS algorithm. To this end, we calculated the proportion of the removed tracklet hypothesis candidates after hypothesis reduction and then counted the number and frequency. Note that an initial feasible solution was necessary in order to reduce the hypothesis candidates, so we applied the method introduced in Sec.VI to generate that. Table IV gives the frequency of five tracklet hypothesis categories.

TABLE V
NUMBER OF DIFFERENT TRACKLET CATEGORIES IN MWIS SOLUTION

Dataset	Tracklet Hypothesis Category				
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$
MOT16	75162 (55.65%)	159 (0.12%)	4097 (3.03%)	55394 (41.02%)	245 (0.18%)
MOT17	149774 (41.68%)	440 (0.12%)	8742 (2.43%)	199691 (55.58%)	653 (0.18%)

In Table IV, 100% means all hypothesis candidates are removed because all of them cannot form a better feasible solution according to our upper bound estimation. Our hypothesis reduction process works the best in this situation; 0%-24% represents less than 25% of candidates that are removed in the hypothesis reduction, while most others still remain. The MWIS problem is not simplified significantly, and our hypothesis reduction process works the least in this situation.

The results in Table IV show that our hypothesis reduction is effective in both datasets: in the MOT16 benchmark, we solved the MWIS problem for 128782 times, and we removed all hypothesis candidates and directly got the optimal solution before solving the NP-hard problem for 126944 times, which is more than 98.5%; in the MOT17 benchmark, the number is 98.3%.

The results of hypothesis reduction prove our initial feasible solution generally works well and the upper bound estimation is tight. Both reasons lead to the effectiveness of hypothesis reduction.

C. Tracklet-Level Association and Iterative MWIS Algorithm

We analyzed the effectiveness of our tracklet-level association and iterative MWIS algorithm. First, we compared the tracking performance of MHT with and without tracklet-level(TA) association as shown in Table VI and Table VII. The ID switches and MOTA become better. It proves that the tracklet-level association can reduce the IDS errors as expected. Then we incorporated iterative MWIS algorithm into MHT on the MOT16 and MOT17 benchmarks, and the results are shown in Table VI and Table VII. We found that the IMWIS improved MOTA in both datasets. It proves that our IMWIS algorithm benefits the tracking performance. Note that we set a dummy detection limitation (< 20) in our method.

The computational time of the traditional MHT tracker on the MOT16 and MOT17 benchmarks are respectively 8141s and 22484s, while the computational time of tracker with our IMWIS are respectively 4980s and 18215s. We found that our iterative MWIS speeded up the tracking process.

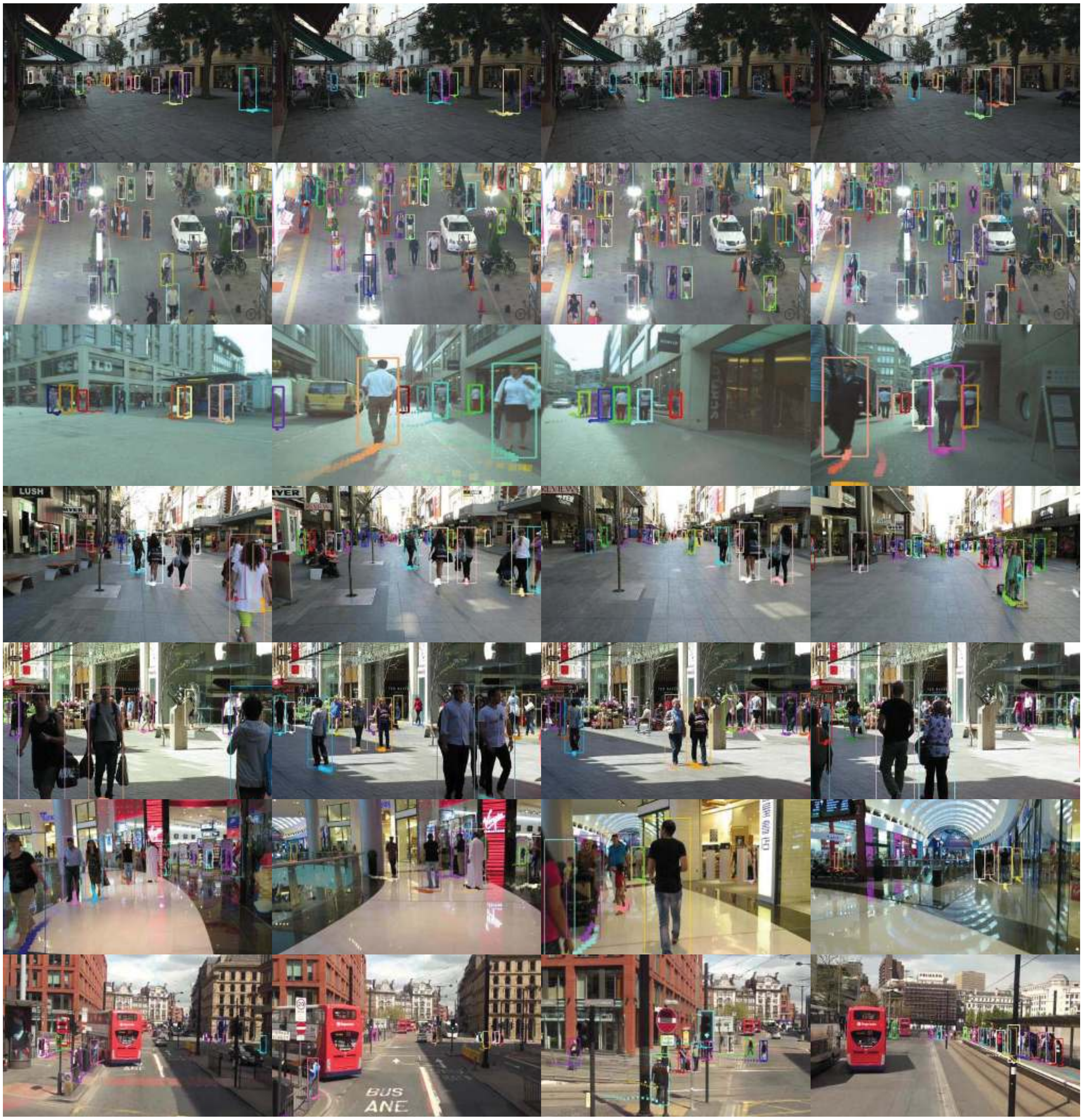


Fig. 5. Tracking results of our method on the MOT16 and MOT17 benchmark. More videos are available on the MOT Challenge website.

TABLE VI

RESULTS ON THE TRAINING SET OF MOT16 BENCHMARK

Method	MOTA↑	MOTP↑	GT	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	Rcll↑	Prcn↑
baseline	39.7	77.8	517	**87**	**234**	5305	60722	511	486	45.0	90.4
TA	40.3	77.9	517	**87**	235	5523	**59934**	420	463	**45.7**	90.0
TA+IMWIS	**41.3**	**78.1**	517	78	237	2626	61818	**408**	**446**	44.0	94.9
Approximation	**41.3**	**78.1**	517	78	**234**	**2510**	61839	430	447	44.0	**95.1**

D. Iterative MWIS Approximation Algorithm

We analyzed our approximation algorithm in this subsection. Our first goal was to prove that our assumption, ignoring the hypotheses in $\mathcal{C}_5$, was reasonable, and to prove that the tracker with approximation algorithm can get nearly the same performance as the original method.

TABLE VII

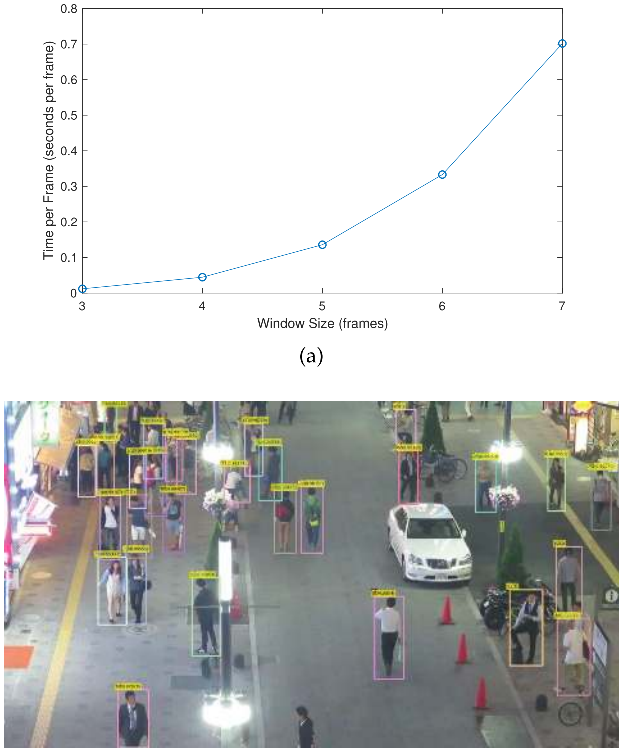
RESULTS ON THE TRAINING SET OF MOT17 BENCHMARK

Method	MOTA↑	MOTP↑	GT	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	Rcll↑	Prcn↑
baseline	51.7	84.6	1638	422	566	11341	149667	1715	1520	55.6	94.3
TA	52.1	84.6	1638	425	561	11575	**148293**	1538	1503	**56.0**	94.3
TA+IMWIS	**52.6**	**84.7**	1638	**435**	556	7834	150236	**1491**	**1485**	55.4	96.0
Approximation	**52.6**	**84.7**	1638	431	**554**	**7655**	150529	1519	1505	55.3	**96.1**

TABLE VIII

THE MOT16 BENCHMARK LEADERBOARD. ACCESSED ON 09/09/2018

Tracker	IDF1↑	MOTA↑	MT↑	ML↓	FP↓	FN↓	ID Sw.↓	Frag↓
NOMT [15]	53.3	46.4	18.3%	41.4%	9,753	87,565	**359**	**504**
MCjoint [14]	52.3	47.1	**20.4%**	46.9%	6,703	89,368	370	598
LMP [16]	51.3	**48.8**	18.2%	40.1%	6,654	**86,245**	481	595
STAM [33]	50.0	46.0	14.6%	43.6%	6,895	91,117	473	1,422
oICF [34]	49.3	43.2	11.3%	48.5%	6,651	96,515	381	1,404
EDMT [25]	47.9	45.3	17.0%	**39.9%**	11,122	87,890	639	946
NLLMPa [17]	47.3	47.6	17.0%	40.4%	**5,844**	89,093	629	768
TLMHT (Ours)	**55.3**	48.7	15.7%	44.5%	6,632	86,504	413	642



(a)

(b)

Fig. 6. Short tracklet generation efficiency. (a) Runtime of the short tracklet generation with different window size, (b) Complex sequence used to calculate runtime.

The number of hypotheses, which are selected in the global hypothesis at each frame is presented in Table V, where the frequency of $\mathcal{C}_5$ is less than 0.2% and therefore proves that our assumption, ignoring the hypotheses in $\mathcal{C}_5$, is a reasonable one.

Moreover, the results of both the iterative MWIS trackers with and without the approximation algorithm on the benchmark of MOT16 and MOT17 are given in Table VI and Table VII. The results show that the tracking performances of both are nearly the same.

The computational time of the original method on the MOT16 and MOT17 benchmarks are respectively $4980s$ and $18215s$, while the time of approximation method on these two datasets are respectively $4567s$ and $16682s$, which are $413s(8.3\%)$ and $1533s(8.4\%)$ faster. Note that from Table IV, we know that even in the original method, more than 98% of MWIS problem is directly solved by our tracklet hypothesis reduction. The less than 2% of MWIS problems contribute to more than 8% faster. The main reason for the runtime improvement is utilizing the solution from previous frame as describe in Alg.2.

The experimental results show our proposed approximation method is practical in terms of both accuracy and efficiency.

E. Benchmark Comparison

Finally, we tested our proposed tracking method on both MOT16 and MOT17 benchmarks, and the quantitative evaluations of our approach, as well as the best previously published approaches, are provided in Table VIII and Table IX. The comparison can also be found in the MOT Challenge website[1]; our tracker is named TLMHT(Tracklet-level Multiple Hypothesis Tracking). Our tracker outperforms all the previously published tracking algorithms.

In all these state-of-the-art trackers, MHT_DAM [7] and EDMT [25] are the best MHT-based tracking methods, which under the same framework with ours. Our tracker outperforms them in both MOT16 and MOT17 benchmarks on MOTA. It is noted that our ID switches and Fragment are significantly less than other methods. The main reason is that tracklets are regarded as tracking unit rather than detections in other MHT methods. Moreover, when we generate short tracklets, the tracker considers affinity of all detection pairs instead of only adjacent detections.

[1] https://motchallenge.net

TABLE IX
THE MOT17 BENCHMARK LEADERBOARD. ACCESSED ON 09/09/2018

Tracker	IDF1↑	MOTA↑	MT↑	ML↓	FP↓	FN↓	ID Sw.↓	Frag↓
EDMT [25]	51.3	50.0	**21.6%**	36.3%	32,279	**247,297**	2,264	3,260
PHD_GSDL [35]	49.6	48.0	17.1%	35.6%	23,199	265,954	3,998	8,886
FWT [36]	47.6	**51.3**	21.4%	**35.2%**	24,101	247,921	2,648	4,279
MHT_DAM [7]	47.2	50.7	20.8%	36.9%	22,875	252,889	2,314	2,865
EAMTT [37]	41.8	42.6	12.7%	42.7%	30,711	288,474	4,488	5,720
TLMHT (Ours)	**56.5**	50.6	17.6%	43.4%	**22,213**	255,030	**1,407**	**2,079**

VIII. CONCLUSION

Multiple hypothesis tracking solves the tracking tasks as a series of maximum weighted independent set problem across the tracking time. Unlike previous works, these NP-hard MWIS problems are solved independently without any prior information of each frame and ignore the relevance between adjacent frames. This paper first defined a new concept of hypothesis category and then presented the category transfer model. By using the model, we presented a novel iterative algorithm that solved the MWIS problem by taking advantages of the previous solution. We also introduced a polynomial-time approximation algorithm to convert the NP-hard problem into a bipartite graph matching problem that can be solved in polynomial time. In addition, we introduced a novel tracklet-level association for multiple hypothesis tracking in order to maintain the tracklet hypothesis number. Our experimental results showed that our iterative algorithm significantly improved the efficiency in solving the MWIS problems in MHT. The tracking performance of our approximate algorithm was quite similar to the original one. We also compared our method with the published state-of-the-art trackers in the benchmark of MOT16 and MOT17, and the overall performance showed that our results were competitive.

REFERENCES

[1] X.-Y. Jing *et al.*, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 695–704.

[2] J. Chen and Q. Ji, "A probabilistic approach to online eye gaze tracking without explicit personal calibration," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1076–1086, Mar. 2015.

[3] F. Zhu, Z. Li, S. Chen, and G. Xiong, "Parallel transportation management and control system and its applications in building smart cities," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1576–1585, Jun. 2016.

[4] S. Zhang, S. Zhao, Y. Sui, and L. Zhang, "Single object tracking with fuzzy least squares support vector machine," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5723–5738, Dec. 2015.

[5] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. AC-24, no. 6, pp. 843–854, Dec. 1979.

[6] D. J. Papageorgiou and M. R. Salpukas, "The maximum weight independent set problem for data association in multiple hypothesis tracking," in *Optimization and Cooperative Control Strategies*. New York, NY, USA: Springer, 2009, pp. 235–255.

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

[8] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Multi-person tracking by multicut and deep matching," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 100–111.

[9] N. Le, A. Heili, and J.-M. Odobez, "Long-term time-sensitive costs for CRF-based tracking by detection," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 43–51.

[10] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1201–1208.

[11] J. Black, T. Ellis, and P. Rosin, "Multi view image surveillance and tracking," in *Proc. Workshop Motion Video Comput.*, Dec. 2002, pp. 169–174.

[12] A. Andriyenko and K. Schindler, "Globally optimal multi-target tracking on a hexagonal lattice," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2010, pp. 466–479.

[13] J. Berclaz, F. Fleuret, and P. Fua, "Multiple object tracking using flow linear programming," in *Proc. 12th IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS-Winter)*, Dec. 2009, pp. 1–8.

[14] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox, and B. Schiele. (2016). "A multi-cut formulation for joint segmentation and tracking of multiple objects." [Online]. Available: https://arxiv.org/abs/1607.06317

[15] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3029–3037.

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

[17] E. Levinkov *et al.*, "Joint graph decomposition & Node labeling: Problem, algorithms, applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1904–1912.

[18] A. Sadeghian, A. Alahi, and S. Savarese. (2017). "Tracking the untrackable: Learning to track multiple cues with long-term dependencies." [Online]. Available: https://arxiv.org/abs/1701.01909

[19] A. Milan, K. Schindler, and S. Roth, "Multi-target tracking by discrete-continuous energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2054–2068, Oct. 2016.

[20] A. A. Butt and R. T. Collins, "Multi-target tracking by Lagrangian relaxation to min-cost network flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1846–1853.

[21] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5537–5545.

[22] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4091–4099.

[23] M. Han, W. Xu, H. Tao, and Y. Gong, "An algorithm for multiple object trajectory tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun./Jul. 2004, p. 1.

[24] M. Manafifard, H. Ebadi, and H. A. Moghaddam, "Appearance-based multiple hypothesis tracking: Application to soccer broadcast videos analysis," *Signal Process., Image Commun.*, vol. 55, pp. 157–170, Jul. 2017.

[25] J. Chen, H. Sheng, Y. Zhang, and Z. Xiong, "Enhancing detection model for multiple hypothesis tracking," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 2143–2152.

[26] W. Brendel, M. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1273–1280.

[27] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 167–172, Jan. 2007.

[28] R. A. Rossi and N. K. Ahmed, "Coloring large complex networks," *Soc. Netw. Anal. Mining*, vol. 4, no. 1, p. 228, Dec. 2014.

[29] S. Cai, "Balance between complexity and quality: Local search for minimum vertex cover in massive graphs," in *Proc. IJCAI*, 2015, pp. 747–753.

[30] S. Cai and J. Lin, "Fast solving maximum weight clique problem in massive graphs," in *Proc. IJCAI*, 2016, pp. 568–574.

[31] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. (2016). "MOT16: A benchmark for multi-object tracking." [Online]. Available: https://arxiv.org/abs/1603.00831

[32] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[33] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu. (2017). "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism." [Online]. Available: https://arxiv.org/abs/1708.02843

[34] H. Kieritz, S. Becker, W. Hübner, and M. Arens, "Online multi-person tracking using integral channel features," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 122–130.

[35] Z. Fu, P. Feng, F. Angelini, J. Chambers, and S. M. Naqvi, "Particle phd filter based multiple human tracking using online group-structured dictionary learning," *IEEE Access*, vol. 6, pp. 14764–14778, 2018.

[36] R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn. (2018). "Fusion of head and full-body detectors for multi-object tracking." [Online]. Available: https://arxiv.org/abs/1705.08314

[37] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, "Online multi-target tracking with strong and weak detections," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 84–99.

[38] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1-1–1-10, Dec. 2008.

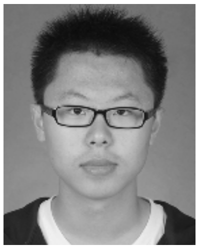
[39] S.-I. Yu, D. Meng, W. Zuo, and A. Hauptmann, "The solution path algorithm for identity-aware multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3871–3879.

[40] A. Maksai, X. Wang, F. Fleuret, and P. Fua, "Non-Markovian globally consistent multi-object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2563–2573.

[41] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 17–35.

Hao Sheng received the B.S. and Ph.D. degrees from the School of Computer Science and Engineering, Beihang University, China, in 2003 and 2009, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, pattern recognition, and machine learning.

Jiahui Chen received the B.S. degree from the School of Advanced Engineering, Beihang University, China, in 2012, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interest is computer vision, and he is particularly interested in multiple object tracking.



Yang Zhang received the B.S. degree from the School of Advanced Engineering, Beihang University, China, in 2014, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interest is computer vision, and he is particularly interested in multiple object tracking.

Wei Ke received the Ph.D. degree from the School of Computer Science and Engineering, Beihang University. He is currently an Associate Professor of computing program with Macao Polytechnic Institute. His research interests include programming languages, image processing, computer graphics, and tool support for object-oriented and component-based engineering and systems. His recent research focuses on the design and implementation of open platforms for the applications of computer graphics and pattern recognition, including programming tools, environments, and frameworks.

Zhang Xiong received the B.S. degree from Harbin Engineering University in 1982 and the M.S. degree from Beihang University, Beijing, China, in 1985. He is currently a Professor with the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, information security, and data vitalization.

Jingyi Yu received the B.S. degree from the California Institute of Technology in 2000 and the Ph.D. degree from MIT in 2005. He is currently a Professor and the Vice Dean of the School of Information Science and Technology, ShanghaiTech University. He was a recipient of the NSF Career Award and currently serves as an Associate Editor of IEEE TPAMI and TIP.