Online Multi-Target Tracking with Unified Handling of Complex Scenarios

Huaizu Jiang, Jinjun Wang, Yihong Gong, *Senior Member, IEEE*
Na Rong, Zhenhua Chai, and Nanning Zheng, *Fellow, IEEE*

***Abstract*—Complex scenarios, including miss detections, occlusions, false detections, and trajectory terminations, make the data association challenging. In this paper, we propose an online tracking-by-detection method to track multiple targets with unified handling of aforementioned complex scenarios, where current detection responses are linked to previous trajectories. We introduce a dummy node to each trajectory to allow it to temporally disappear. If a trajectory fails to find its matching detection, it will be linked to its corresponding dummy node until the emergence of its matching detection. Source nodes are also incorporated to account for the entrance of new targets. The standard Hungarian algorithm, extended by the dummy nodes, can be exploited to solve the online data association implicitly in a global manner, although it is formulated between two consecutive frames. Moreover, as dummy nodes tend to accumulate in a fake or disappeared trajectory while they only occasionally appear in a real trajectory, we can deal with false detections and trajectory terminations by simply checking the number of consecutive dummy nodes. Our approach works on a single, uncalibrated camera, and requires neither scene prior knowledge nor explicit occlusion reasoning, running at 132 frame per second (fps) on the PETS09-S2L1 benchmark sequence. Experimental results validate the effectiveness of the dummy nodes in complex scenarios and show that our proposed approach is robust against false detections and miss detections. Quantitative comparisons with other methods on five benchmark sequences demonstrate that we can achieve comparable results with most existing offline methods and better results than other online algorithms.**

***Index Terms*—Multi-target tracking, complex scenarios**

I. Introduction

Multi-target tracking has a wide range of applications in the areas of video surveillance, robotics, video content understanding, etc. With rapid improvements of object detectors (*e.g.*, HOG [1] and DPM [2]), tracking-by-detection methods have received a lot of research interests in recent years, where the tracking task is accomplished by finding correspondences among detection responses in different frames or temporal windows of a video to form a set of coherent trajectories. This is usually called *data association*.

In this paper, we study the data association problem for *online* multi-target tracking through a single uncalibrated camera. When dealing with real-world data, *complex scenarios* must be handled to achieve appealing tracking results. On one hand, detection failures are inevitable, which include the miss detection where a target is misclassified as the background, the false detection where a background region is incorrectly recognized as a target, and the occlusion where an object is partially or fully invisible because of the limited camera view. Therefore, there exists misalignment between trajectories and detections during the data association where not every trajectory or detection can find its correspondence. On the other hand, in real-world scenarios, targets may appear and disappear anytime and anywhere in the scene. We need to automatically tackle the initializations and terminations of trajectories to accommodate dynamic target changes. All these complex scenarios make the data association challenging.

To deal with such complex scenarios, various algorithms have been proposed in the past decade. Detection failures are addressed by the continuous confidence output along with Particle Filter [3] and the explicit occlusion reasoning [4]. Alternatively, they can be addressed in a global temporal window using the network flow [5], high-order energy minimization [6], [7], and hierarchical data association [8], [9]. Furthermore, it is often assumed that new targets may enter or move out of the scene from some certain areas, *e.g.*, the border of the camera view, in order to deal with trajectory initializations and terminations. However, these data association models are either difficult to estimate (*e.g.*, explicit occlusion reasoning) from a single camera view [4], [5] or subject to high computational burdens [3], [6], [7], [8], [9], [10], restricting their adaptations to time-critical scenarios (*e.g.*, surveillance). Additional reliance on the scene prior knowledge may also limit their adaptability to wider application areas.

In this paper, we propose an online multi-target tracking method with unified handling of aforementioned complex scenarios. During the data association, current detection responses are gradually linked to existing trajectories. We introduce a *dummy node* for each trajectory to explicitly handle miss detections and occlusions, giving each trajectory opportunities to temporally disappear. If a trajectory fails to find its matching detection, it will be linked to its corresponding dummy node until its matching detection resurfaces. To account for the trajectory initialization, we also introduce a *source node* for each detection. A new target will be automatically linked to its corresponding source node and a new trajectory will be created if it does not match any existing trajectories. Data association can therefore be efficiently solved using the Hungarian algorithm [11] in polynomial time on a balanced bipartite graph. Although formulated between each two consecutive frames for online multi-target tracking, our data association can implicitly span multiple frames using dummy nodes. Another advantage

Huaizu Jiang, Jinjun Wang, Yihong Gong, Na Rong, and Nanning Zheng are with Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

Zhenhua Chai is with Media Technology Lab, Central Research Institute at Huawei Technologies Co.,Ltd, Beijing, China.

Corresponding author: Yihong Gong. Email: ygong@mail.xjtu.edu.cn.

of the dummy node is its potential to deal with false detections and trajectory terminations. Since false detections are random and unstable, a fake trajectory initialized with a false detection can not find consistent correspondences in subsequent frames and thus a set of dummy nodes will accumulate during the tracking process, making it easy to be distinguished from a true one. Similarly, dummy nodes will also accumulate in a disappeared trajectory. We can uniformly deal with false detections and trajectory terminations by simply checking the accumulated number of dummy nodes. Though incorporating scene prior knowledge, *e.g.*, the entrance and exit area of trajectories, might be beneficial, our proposed approach does not necessarily rely on it.

Our contributions therefore include three fold: 1) our approach is capable of tracking multiple targets with unified handling of complex scenarios, which does not necessarily reply on any scene-specific prior knowledge nor explicit occlusion reasoning; 2) we propose dummy nodes and source nodes for online data association, extending the standard Hungarian algorithm; 3) our proposed online multi-target tracking method is fast and efficient, running at 132 frames per second (fps) on the PETS09-S2l1 benchmark sequence. Experiments on benchmark datasets validate the effectiveness of our approach, where we can achieve comparable results with most existing offline methods and better results than other online algorithms.

The rest of the paper is organized as follows. Section II introduces related works and discusses their differences from our proposed approach. Our proposed method is presented in Section III. Section IV demonstrates experimental results, where we examine the effectiveness of dummy nodes and robustness of our approach against false detections and miss detections. Comparisons with state-of-the-art approaches are also presented. Section V concludes this paper.

II. Related Work

There is an extensive research literature on multi-target tracking. In this section, we briefly discuss offline and online tracking-by-detection approaches proposed in recent years.

Offline multi-target tracking. Offline methods solve the data association problem in a large temporal window or even the whole sequence. In [5], a method is proposed to model the multi-target tracking as a maximum-a-posteriori problem (MAP) and trajectories are found by solving the min-cost flow in the cost-flow network. Occlusions are handled by iteratively adding occluded object hypothesis into the network based on explicit occlusion reasoning. This method is extended in [12] to find the near-optimal solution, where each trajectory is sequentially recovered based on dynamic programming. High-order track smoothness constraints (*i.e.*, constant-velocity in a path) are incorporated into the cost-flow network formulation and are iteratively relaxed using Lagrangian relaxation [13].

With the hierarchical association models proposed in [8], [14], detection responses are first heuristically linked into tracklets (*i.e.*, short trajectories). Affinity measures are then computed between tracklets to find optimal matchings among all tracklets in a large temporal window using the standard Hungarian algorithm [11]. Miss detections and occlusions are dealt with by linking two non-consecutive tracklets. In [14], the merge-split measurements are embedded into the optimal matching problem to better preserve the object identity. In a recent work [10], trajectories are iteratively obtained by finding the Generalized Minimum Clique Graph (GMCP) based on a hierarchical association framework, where each node corresponds to a detection response or a tracklet. Hypothetical nodes, whose appearance and motion attributes need to be explicitly computed based on the inliers/outliers estimations for each tracklet/trajectory, are iteratively incorporated into the graph to deal with miss detections and occlusions. In a latest work [15], data association of tracklets in a global temporal window is defined as a Generalized Maximum Multi Clique problem (GMMCP). To deal with miss detections and occlusions, dummy nodes are inserted for each clique. Moreover, aggregated dummy nodes are introduced to speed up the data association.

To further achieve better performance, data-driven approaches are also proposed to leverage machine learning algorithms to find the correspondences of detections or tracklets. In [16], affinity of each candidate correspondence pair is learned based on appearance features using the boosting algorithm. In addition to the appearance, motion patterns can also be online learned to achieve more robust performance [17]. In [9], the correspondence between two tracklets is modeled as a random variable. An online learned Conditional Random Field is exploited to find the optimal solution by considering pairwise constraints of correspondences.

Minimizing high-order energy functions to find optimal trajectories is another popular research direction. A continuous energy is introduced in [6] to find the optimal trajectories by imposing several constraints on the candidate solutions (*e.g.*, mutual exclusion and target persistence). The energy function is then extended to the discrete-continuous domain [7] to simultaneously solve data association and trajectory fitting. In a recent work [18], more sophisticated constraints based on statistical properties of real trajectories are incorporated. In these approaches, miss detections and occlusions are implicitly addressed by interpolation during the fitting of trajectories.

Online multi-target tracking. Alternatively, Particle Filter is adopted for tracking multiple persons [3] only based on past frames. To avoid miss detections, the continuous confidence output is adopted. Since in real-world scenarios, people always interact with others and the environment, there also exist approaches that model social behaviors of pedestrians for more robust tracking, especially when occlusions exist [19], [20]. Recently, two online multi-target tracking approaches [21], [4] are presented. Wu *et al.* [21] propose to represent each detection by multiple patches, whose motion directions are estimated locally. The correspondence between a trajectory and a detection is estimated by examining the agreement of the global motion of the trajectory and the local motion of the detection. Possegger *et al.* [4] instead exploit the geometric information to find correspondences of trajectories and visible detections based on explicit occlusion reasoning and estimations of occlusion geodesics.

Our proposed method falls in the online multi-target tracking category. Compared with offline algorithms, our approach

Fig. 1. Illustration of effectiveness of dummy nodes and source nodes for data association in online multi-target tracking. **Top row**: a dummy node (denoted with the dashed bounding box with magenta color) is inserted in frame #57 (middle column) to account for the occlusion. **Bottom row**: with the help of both dummy node and source node, we can automatically discover the trajectory #7 (brown color) in frame #15 (middle column) when the woman enters the scene.

does not rely on future frames. Augmented with dummy nodes, our online data association can span multiple frames and be implicitly solved in a global manner. Therefore we can still achieve comparable results with most existing offline methods. Moreover, our dummy node is quite different from the hypothetical node of GMCP tracker [10] and dummy node of GMMCP tracker [15]. First of all, the hypothetical node of GMCP tracker and dummy node of GMMCP tracker are only designed to deal with miss detections and occlusions, while our approach is capable of handling complex scenarios, including miss detections, occlusions, false detections, trajectory initializations and terminations, in a unified way with the help of dummy nodes. Second, both the hypothetical node and dummy node in the above works are inserted in a global temporal window for *offline* multi-target tracking algorithms, while our dummy node is incorporated into the data association for *online* multi-target tracking. Finally, appearance and motion features of a hypothetic node need to be computed based on the inlier/outlier estimations in a global temporal window, while our dummy node serves as an indicator of the invisibility of a trajectory only, leading to an efficient multi-target tracking algorithm.

Compared with online models, deploying social behavior model for multi-target tracking requires additional scene knowledge, *e.g.*, ground plane estimation [19], annotations of obstacles and social groups [20], which can not be offered by a single uncalibrated camera. Among others, the most similar models to ours might be [21] and [4], which solve the data association problem between each two consecutive frames using the Hungarian algorithm [11] as well. Our proposed method differs from them on two aspects. On one hand, we require neither local motion estimation that might be time consuming nor explicit occlusion reasoning that might be ambiguous through a single camera view. On the other hand, we extend the standard Hungarian algorithm with dummy nodes, which offer benefits to handle not only miss detections and occlusions explicitly but also false detections and trajectory terminations cheaply. Moreover, compared with traditional particle filter-based methods, our approach does not suffer from exponentially increasing complexity.

III. Proposed Method

In this section, we present our proposed method. The online data association for multi-target tracking is presented in Section III-A with the introduction of dummy nodes and source nodes. Observation model is given in Section III-B. Updating of observation models is presented in Section III-C. We then demonstrate how to deal with complex scenarios including miss detections, occlusions, false detections, and trajectory terminations in a unified manner in Section III-D. Finally, Section III-E analyzes the time complexity of our proposed approach.

A. Online Data Association

Given the trajectories $\mathcal{T}^t$ in the frame I^t and detections $\mathcal{X}^{t+1}$ in the frame I^{t+1}, our goal is to find correspondences between $\mathcal{T}^t$ and $\mathcal{X}^{t+1}$. Misalignment exists between them where not every trajectory or detection can find its matching, making the data association complicated. On one hand, due to limitations of existing object detectors, miss detections are inevitable in practice. Occlusions often occur as well. For example, in Fig. 1, there is no matching for trajectory #5 (with magenta color) in frame #57 because of the occlusion. Its matching detection re-surfaces in frame #63 and the subsequent frames. This example suggests that it is necessary to examine more than two neighboring frames to find correct correspondences between trajectories and detections. Considering more than two consecutive frames, however, requires future frames and will lead to more computational burden. On the other hand, a new target may enter the scene at any time. For instance, a new target (with orange color) enters the scene in frame #15 shown in the second row of Fig. 1, which can not find its corresponding trajectory during the data association. In addition, the existence of false detections makes the automatic discovery of new trajectories much harder. Therefore, we introduce dummy nodes and source nodes to deal with the misalignment.

Assume that there are J_t trajectories $\mathcal{T}^t = \{T_i^t\}_{i=1}^{J_t}$ in the frame I^t and K_{t+1} detections $\mathcal{X}^{t+1} = \{x_j^{t+1}\}_{j=1}^{K_{t+1}}$ in the frame I^{t+1}, where T_i^t denotes the i'th trajectory in I^t and x_j^{t+1} the j'th detection in I^{t+1}. To address the misalignment between $\mathcal{T}^t$ and $\mathcal{X}^{t+1}$, we introduce J_t *dummy nodes* $\mathcal{D}^{t+1} = \{d_i^{t+1}\}_{i=1}^{J_t}$ in the frame I^{t+1} for each trajectory in $\mathcal{T}^t$. If the trajectory T_i^t fails to find its optimal matching to any of the detections in $\mathcal{X}^{t+1}$ because of either miss detection or occlusion, it will be linked to its corresponding dummy node d_i^{t+1}. The dummy node d_i^{t+1} allows the trajectory T_i^t to be temporally invisible, enabling the association to span multiple frames. For instance, as shown in the first row of Fig. 1, the trajectory #5 links to the dummy node (denoted with dashed the bounding box) due to occlusions from frame #57 to frame #62.

Similarly, a new target in $\mathcal{X}^{t+1}$ may not find its matching trajectory in $\mathcal{T}^t$ either. Therefore we also introduce K_{t+1} *source nodes* $\mathcal{S}^t = \{s_j^t\}_{j=1}^{K_{t+1}}$ into frame I^t where s_j^t corresponds to x_j^{t+1}. x_j^{t+1} will be linked to s_j^t and a new trajectory is created if there is no matching trajectory in $\mathcal{T}^t$ for x_j^{t+1}.

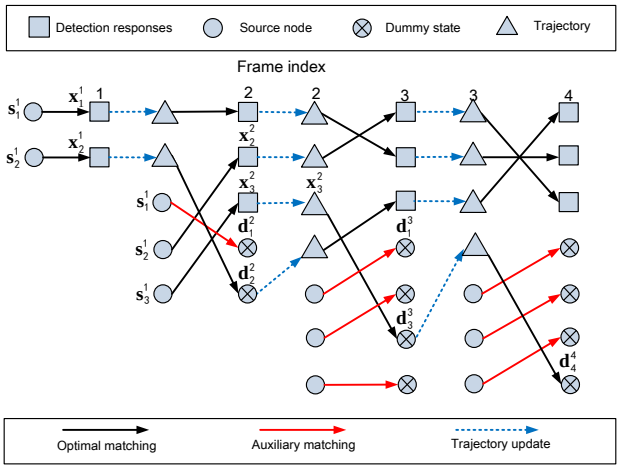

Fig. 2. Illustration of data association for online multi-target tracking. Three source nodes are created in frame I^1 since there are three detection responses in frame I^2. Similarly two dummy nodes are created in frame I^2 since there are two trajectories in frame I^1. After optimization of Eq(2) between I^1 and I^2, 4 optimal matchings and an auxiliary matching are generated. The auxiliary matching is deleted and trajectories are updated according to the optimal matchings.

An illustration of dummy nodes and source nodes is shown in Fig. 2.

With the introduction of dummy nodes and source nodes, we align trajectories and detections and get a balanced bipartite graph $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ where $\mathcal{U} = \{u_i\}_{i=1}^{J_t+K_{t+1}}$, $u_i \in \mathcal{T}^t \cup \mathcal{S}^t$, and $\mathcal{V} = \{v_j\}_{j=1}^{J_t+K_{t+1}}$, $v_j \in \mathcal{X}^{t+1} \cup \mathcal{D}^{t+1}$. Without loss of generality, we assume vertices between $\mathcal{U}$ and $\mathcal{V}$ are fully connected. The data association is then to find the maximum weighted matchings $\mathbf{M} = [m_{ij}]$ on the bipartite graph $\mathcal{G}$, where $m_{ij} \in \{0, 1\}$ denotes if u_i matches v_j (1 for matching).

The matching matrix $\mathbf{M}$ can be divided into four parts:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{tx}^t & \mathbf{M}_{td}^t \\ \mathbf{M}_{sx}^t & \mathbf{M}_{sd}^t \end{bmatrix}, \tag{1}$$

where the first block matrix $\mathbf{M}_{tx}^t$ denotes the matching matrix between trajectories $\mathcal{T}^t$ and detections $\mathcal{X}^{t+1}$. The second part $\mathbf{M}_{td}^t$, which explicitly deals with miss detections and occlusions, corresponds to the matching matrix between trajectories $\mathcal{T}^t$ and dummy nodes $\mathcal{D}^{t+1}$. The third one $\mathbf{M}_{sx}^t$, which can automatically discover new trajectories, denotes the matching matrix between source nodes $\mathcal{S}^t$ and detections $\mathcal{X}^{t+1}$. The last one $\mathbf{M}_{sd}^t$ serves as a place holder only.

Given the similarity matrix $\mathbf{W} = [w_{ij}]$ between $\mathcal{U}$ and $\mathcal{V}$, where $w_{ij} \in \mathbb{R}^+$ captures the similarity between u_i and v_j, the optimal data association can be achieved by solving the following optimization problem

$$\begin{aligned} \mathbf{M}^* &= \arg\max_{\mathbf{M}} \sum_{i,j} w_{ij} m_{ij}, \\ s.t. \quad & \sum_{j=1}^{J_t+K_{t+1}} m_{ij} = 1, \forall i, \\ & \sum_{i=1}^{J_t+K_{t+1}} m_{ij} = 1, \forall j, \end{aligned} \tag{2}$$

where the constraints of m_{ij} reveal the *mutual exclusion* fact that two trajectories will not simultaneously occupy the same detection. Such an optimization problem can be efficiently solved using the Hungarian algorithm [11].

B. Observation Model

In this section, we describe how to compute the similarity matrix $\mathbf{W}$ given the observations. Similar to the matrching matrix $\mathbf{M}$, the similarity matrix $\mathbf{W}$ can also be divided into four parts,

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{tx}^t & \mathbf{W}_{td}^t \\ \mathbf{W}_{sx}^t & \mathbf{W}_{sd}^t \end{bmatrix}. \tag{3}$$

Similarity between $\mathcal{T}^t$ and $\mathcal{X}^{t+1}$. $\mathbf{W}_{tx}^t = \left[c(T_i^t, x_j^{t+1})\right]$ defines similarity scores between trajectories $\mathcal{T}^t$ and detections $\mathcal{X}^{t+1}$. $c(T_i^t, x_j^{t+1})$ is the similarity score between T_i^t and x_j^{t+1}.

In this paper, we adopt the bounding box representation, and hence each detection x_i^t is represented as $x_i^t = (\mathbf{p}_i^t, \mathbf{q}_i^t, \mathbf{c}_i^t)$, where $\mathbf{p}_i^t, \mathbf{q}_i^t$ are the central position and the size (width, height) of the bounding box, respectively. $\mathbf{c}_i^t$ is the HSV histogram of the bounding box area with 10 bins for each channel.

Similarly, each trajectory T_i^t is represented as a set of attributes $T_i^t = (\mathcal{P}_i^t, \mathcal{Q}_i^t, \mathbf{a}_i^t, L_i^t, A_i^t, B_i^t)$, where $\mathcal{P}_i^t$ is the set of central positions of the detections that form the trajectory T_i^t, and can be expressed as $\mathcal{P}_i^t = \{\mathbf{p}_u^v \,|\, x_u^v \in T_i^t, 1 \le v \le t\}$; $\mathcal{Q}_i^t$ is the set of dimensions of the detections forming the trajectory, and can be written as $\mathcal{Q}_i^t = \{\mathbf{q}_u^v \,|\, x_u^v \in T_i^t, 1 \le v \le t\}$; $\mathbf{a}_i^t$ is the appearance descriptor defined as a time-weighted average of the color histograms of the detections $x_u^v \in T_i^t$; L_i^t is the number of detection responses forming T_i^t; A_i^t and B_i^t indicate the starting frame number and the number of consecutive dummy nodes, respectively.

The similarity score $c(T_i^t, x_j^{t+1})$ is then defined as follows,

$$c(T_i^t, x_j^{t+1}) = -\log\left(1 - \Phi(T_i^t, x_j^{t+1})\right), \tag{4}$$

where $\Phi(T_i^t, x_j^{t+1})$ is further defined as

$$\begin{aligned} \Phi(T_i^t, x_j^{t+1}) = \phi_p(T_i^t) \big(& w_m \phi_m(T_i^t, x_j^{t+1}) + \\ & w_s \phi_s(T_i^t, x_j^{t+1}) + w_a \phi_a(T_i^t, x_j^{t+1})\big). \end{aligned} \tag{5}$$

In the above equation, $\phi_p(T_i^t)$ measures the persistence of T_i^t, defined as:

$$\phi_p(T_i^t) = \frac{1}{1 + e^{-L_i^t}}. \tag{6}$$

It is clear that $\phi_p(T_i^t)$ gives high scores to longer trajectories and penalizes shorter ones. $\phi_m(T_i^t, x_j^{t+1})$, $\phi_s(T_i^t, x_j^{t+1})$, and $\phi_a(T_i^t, x_j^{t+1})$ compute the similarity score between T_i^t and x_j^{t+1} in terms of motion trend, size, and appearance cues respectively. w_m, w_s and w_a are their weights ($w_a = w_s = 0.3, w_m = 0.4$ in our implementation). These functions are described as follows.

For the motion trend cue, we check how close the central position of x_j^{t+1} is to the predicted position of T_i^t in the frame I^{t+1},

$$\phi_m(T_i^t, x_j^{t+1}) = \mathrm{G}(\hat{\mathbf{p}}_i^{t+1} - \mathbf{p}_j^{t+1}, \mathbf{\Sigma}_m), \tag{7}$$

where $\mathrm{G}(\cdot, \Sigma)$ is a zero-mean Gaussian function, and $\mathbf{\Sigma}_m$ is a diagonal covariance matrix for the motion similarity. $\hat{\mathbf{p}}_i^{t+1} = g(\mathcal{P}_i^t, t+1, R; \mathbf{\Theta}, \mathbf{H})$ is the predicted central position of T_i^t in I^{t+1} using its latest R detections based on the extended Kalman filter [22], where $\mathbf{\Theta}, \mathbf{H}$ are the state and observation matrices, respectively. In our implementation, we set $R = \min(L_i^t, 10)$ and

$$\mathbf{\Theta} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \quad (8)$$

In real-world scenarios, targets always move with certain physical constraints (*e.g.*, limited velocity). Therefore we make a restriction that

$$c(T_i^t, x_j^{t+1}) = 0 \quad \text{if } \phi_m(T_i^t, x_j^{t+1}) < \tau, \quad (9)$$

where τ is set to 0.25 in our implementation. This prevents the case that one target suddenly moves from one place to a distant one at a high speed.

Similarly, for the size similarity, we check the difference between the dimension of x_j^{t+1} and the predicted dimension of T_i^t

$$\phi_s(T_i^t, x_j^{t+1}) = G(\hat{\mathbf{q}}_i^{t+1} - \mathbf{q}_j^{t+1}, \mathbf{\Sigma}_s), \quad (10)$$

where $\hat{\mathbf{q}}_i^{t+1} = g(\mathcal{Q}_i^t, t+1, J; \mathbf{\Theta}, \mathbf{H})$. $\mathbf{\Sigma}_s$ is a diagonal covariance matrix for the size similarity.

For the appearance cue, we simply check the similarity of their appearance descriptors (*i.e.*, the color histograms),

$$\phi_a(T_i^t, x_j^{t+1}) = \chi^2(\mathbf{a}_i^t, \mathbf{c}_j^{t+1}). \quad (11)$$

Similarity between T^t and $\mathcal{D}^{t+1}$. $\mathbf{W}_{td}^t$ is a *diagonal* matrix of size $J_t \times J_t$ defining similarity scores between trajectories $\mathcal{T}^t$ and dummy nodes $\mathcal{D}^{t+1}$. Here, we adopt fixed similarity scores for this type of connections, indicating that each trajectory has a uniform priori probability to be temporally invisible.

$$\mathbf{W}_{td}^t = \begin{bmatrix} -\log(1-\tau) & & \\ & \ddots & \\ & & -\log(1-\tau) \end{bmatrix}. \quad (12)$$

Similarity between $\mathcal{S}^t$ and $\mathcal{X}^{t+1}$. $\mathbf{W}_{sx}^t$ is a *diagonal* matrix of size $K_{t+1} \times K_{t+1}$ that defines similarity scores between source nodes $\mathcal{S}^t$ and detection responses $\mathcal{X}^{t+1}$ as follows:

$$c(s_j^t, x_j^{t+1}) = -\log\left(1 - \Phi(s_j^t, x_j^{t+1})\right), \quad (13)$$

where $\Phi(s_j^t, x_j^{t+1}) = \max_{T_i^t}\left(\Phi(T_i^t, x_j^{t+1})\right)$. This definition encourages linking a detection response x_j^{t+1} to its source node s_j^t if there is no strong enough connection to it from any trajectories in $\mathcal{T}^t$. Hence

$$\mathbf{W}_{sx}^t = \begin{bmatrix} c(s_1^t, x_1^{t+1}) & & \\ & \ddots & \\ & & c(s_{K_{t+1}}^t, x_{K_{t+1}}^{t+1}) \end{bmatrix}. \quad (14)$$

Similarity between $\mathcal{S}^t$ and $\mathcal{D}^{t+1}$. $\mathbf{W}_{sd}^t$ is used to generate an matching to link the source node s_i^t to the dummy node d_j^{t+1} if there are neither entrances of new targets nor miss detections or occlusions for existing trajectories. Without any prior knowledge of trajectories and detections, $\mathbf{W}_{sd}^t$ is defined using a uniform priori similarity score,

$$\mathbf{W}_{sd}^t = \begin{bmatrix} -\log(1-\tau) & \cdots & -\log(1-\tau) \\ \vdots & \ddots & \vdots \\ -\log(1-\tau) & \cdots & -\log(1-\tau) \end{bmatrix}, \quad (15)$$

with dimension of $J_t \times K_{t+1}$.

C. Updating Observation Models

After solving the maximum weighted matching $\mathbf{M}^*$ on the bipartite graph formed between I^t and I^{t+1}, we get two types of matchings demonstrated with dashed arrows in Fig. 2. Miss detections and occlusions can be handled in optimal matchings, while auxiliary matchings are only place holders. Apparently, there are no needs to create new trajectories for them.

According to the optimal matchings, we should update the observation models of trajectories to take the latest temporal information into consideration. If x_j^{t+1} is connected to T_i^t, the corresponding trajectory T_j^{t+1} is updated as follows,

$$\begin{aligned} \mathcal{P}_j^{t+1} &= \mathcal{P}_i^t \cup \mathbf{p}_j^{t+1}, \\ \mathcal{Q}_j^{t+1} &= \mathcal{Q}_i^t \cup \mathbf{q}_j^{t+1}, \\ \mathbf{a}_j^{t+1} &= \alpha \mathbf{a}_i^t + (1-\alpha)\mathbf{c}_j^{t+1}, \\ L_j^{t+1} &= L_i^t + 1, \\ B_j^{t+1} &= 0, \end{aligned} \quad (16)$$

where α is a time factor with value less than one (set to 0.5).

If a dummy node d_j^{t+1} is linked to T_i^t, suppose T_i^t will be updated to T_j^{t+1} in the frame I^{t+1}, updating T_j^{t+1} is then as follows,

$$\begin{aligned} \mathcal{P}_j^{t+1} &= \mathcal{P}_i^t, \\ \mathcal{Q}_j^{t+1} &= \mathcal{Q}_i^t, \\ \mathbf{a}_j^{t+1} &= \mathbf{a}_i^t, \\ L_j^{t+1} &= L_i^t, \\ B_j^{t+1} &= B_i^t + 1. \end{aligned} \quad (17)$$

D. Unified Handling of Complex Scenarios

When a new trajectory T_i^t is created with the help of a source node in the frame I^t, we have no idea to immediately tell whether it is initialized with a true or false detection. Thus we postpone such a decision by putting T_i^t into a maintained set of *new* trajectories $\mathcal{T}_{new}^t$. When strong enough evidence is available, we either move T_i^t to *active* trajectories $\mathcal{T}_{act}^t$ if it is believed to be indeed a true trajectory or to *fake* trajectories $\mathcal{T}_{fal}^t$ if it is initialized by a false detection. To account for terminations of trajectories, *e.g.*, the trajectory T_i^t may move out of the scene and is not available for future data association. We additionally maintain a set of *dead* trajectories $\mathcal{T}_{dea}^t$. Clearly, only new and active trajectories have the chance to find their correspondences to the detections. Therefore, the data association is solved between $\mathcal{T}_{act}^t \cup \mathcal{T}_{new}^t$ and $\mathcal{X}^{t+1}$ as

Algorithm 1 TrajectoriesUpdating

Input: $\mathcal{T}_{act}^{t}, \mathcal{T}_{dea}^{t}, \mathcal{T}_{new}^{t}, \mathcal{T}_{fal}^{t}$, $\mathbf{M}^{*}$
Output: $\mathcal{T}_{act}^{t+1}, \mathcal{T}_{dea}^{t+1}, \mathcal{T}_{new}^{t+1}, \mathcal{T}_{fal}^{t+1}$
$\mathcal{T}_{act}^{t+1} = \emptyset, \mathcal{T}_{dea}^{t+1} = \emptyset, \mathcal{T}_{new}^{t+1} = \emptyset, \mathcal{T}_{fal}^{t+1} = \emptyset$
for $T_i^t \in \mathcal{T}_{new}^t$ **do**
  Update T_i^t to T_j^{t+1} according to Eq.(16) and Eq.(17)
  if $L_j^{t+1} < t_1, B_j^{t+1} > t_2$ **then**
    insert T_j^{t+1} to $\mathcal{T}_{fal}^{t+1}$ (conversion C1)
  else if $L_j^{t+1} > t_3, B_j^{t+1} = 0$ **then**
    insert T_j^{t+1} to $\mathcal{T}_{act}^{t+1}$ (conversion C2)
  else
    insert T_j^{t+1} to $\mathcal{T}_{new}^{t+1}$ (conversion C3)
  end if
end for
for $T_i^t \in \mathcal{T}_{act}^t$ **do**
  Update T_i^t to T_j^{t+1} according to Eq.(16) and Eq.(17)
  if $B_j^{t+1} > t_4$ **then**
    insert T_j^{t+1} to $\mathcal{T}_{dea}^{t+1}$ (conversion C4)
  else
    insert T_j^{t+1} to $\mathcal{T}_{act}^{t+1}$ (conversion C5)
  end if
end for

introduced in Section III-A, *i.e.*, $\mathcal{T}^t = \mathcal{T}_{act}^t \cup \mathcal{T}_{new}^t$. Initially, all the trajectories in the first frame are in the new trajectories $\mathcal{T}_{new}^1$.

These four types of trajectories will be updated during the tracking process, where complex scenarios will be handled in a unified manner. Detailed trajectories updating is presented in Algorithm 1. Miss detections and occlusions are addressed in conversions C2 and C5. Once a new trajectory has sufficient detections, we believe it is a true trajectory and move it to the active trajectory set when there is no miss detection or occlusion in its latest frame (thus no dummy node in the trajectory. See the conversion C2). During the data association, the mechanism of dummy nodes allows a trajectory to be temporally invisible. Therefore, we consider a trajectory as active if the number of accumulated dummy nodes is under a certain level (see the conversion C5).

Furthermore, our proposed dummy nodes are also capable of handling false detections and trajectory terminations. The conversion C1 deals with false detections and the conversion C4 handles trajectories terminations. Our main insight of explicitly distinguishing false detections from true ones is that false detections are generally unstable, random, and inconsistent compared to true detections [2]. Therefore fake trajectories have the tendency to link to dummy nodes in the future. Similarly when an active trajectory terminates, a set of dummy nodes will accumulate in it as well. This implies that we can uniformly cope with false detections and trajectory terminations by simply checking the number of *successive* dummy nodes.

Although the standard Hungarian algorithm can also deal with the unbalanced bipartite graph matching by inserting miscellaneous rows or columns, we extend it by adding dummy nodes and source nodes to explicitly account for complex scenarios. More importantly, this method has additional benefits to cheaply handle false detections and trajectory terminations in a unified manner.

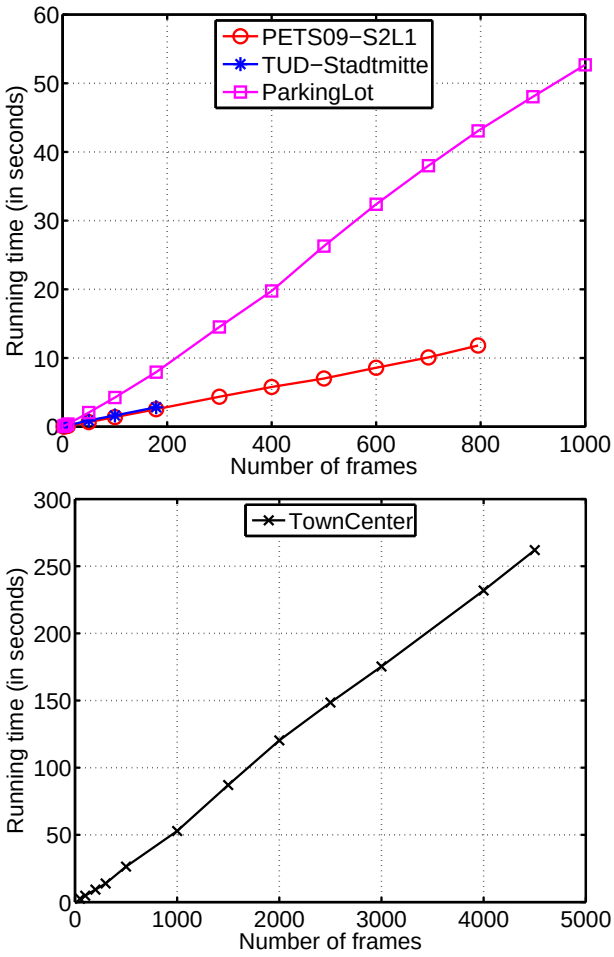

Fig. 3. Running time of our approach on four benchmark sequences. The time complexity of our proposed approach is dependent on the specific scenario, and stable during the tracking process.

Unlike many existing approaches, when dealing with miss detections and occlusions, our proposed method does not require time consuming occlusion reasoning [5], [4], which might be ambiguous and unreliable especially from a single camera view. It is also worth pointing out that our approach does not necessarily rely on the assumption of scene prior knowledge to handle false detections and trajectory terminations. Some algorithms [3], [4] assume that a trajectory usually enters and moves out of the scene from some certain areas, *e.g.*, the border of the camera view. Trajectories starting out of these pre-defined areas are all considered to be fake. Though better performance might be achieved by considering these scene priors, our method works well in practice by simply checking the successive number of dummy nodes, implying that it is more robust in general cases and more suitable for time-critical scenarios.

E. Time Complexity

We adopt the Hungarian algorithm (also know as the Kuhn-Munkres algorithm) to efficiently solve Eq.(2) in polynomial

time. It runs in $O(K^3)$ in the *worst case*, where $K = J_t + K_{t+1}$. Luckily, based on our observation model introduced in Section III-B, the similarity matrix $\mathbf{W}$ is always sparse. In experiments, we observe that it takes around $O(K^2)$ to solve Eq.(2) in the *average case*. Therefore, our proposed method runs in $O(NK^2)$ to track multiple targets in the average case. Moreover, dummy nodes provide a computationally cheap mechanism to deal with false detections and trajectory terminations in a unified manner.

Fig. 3 presents the running time (given the detection output[1]) of our method against the number of video frames on four benchmark sequences. As we can see, the running time behaves almost linearly w.r.t the number of video frames. It reveals the fact that our approach is suitable for time-critical scenarios like the video surveillance.

IV. Experiments

In this section, we analyze our approach in terms of the effectiveness of dummy nodes, robustness, and running time on five benchmark and make comparisons with state-of-the-art offline and online multi-target tracking methods.

A. Dataset

We evaluate the proposed method on five benchmark sequences to test its adaptability to different scenarios (*e.g.*, different frame rates, resolution, time span, and crowdedness).

PETS09-S2L1 [24]: This sequence contains 795 frames. The main challenge is that some targets move with highly non-linear patterns. Occlusions of targets are frequent. We adopt the detections and ground truth provided by [17] in our experiments.

PETS09-S2L2 [24]: It consists of 465 frames. Since people walk very closely in such a crowded scene, there exist heavy occlusions, making it extremely challenging. We adopt the detections and ground truth annotations provided by [25].

TUD-Stadtmitte [26]: There are 179 frames in this sequence. The camera angle is relatively low and there are heavy mutual occlusions, sometimes even full occlusions between targets. We adopt the object detector [2] to generate detections and ground truth annotations provided by [9].

ParkingLot [10]: This is introduced for multi-target tracking evaluations recently. There are 1000 frames with moving crowded targets. Mutual occlusions of targets are much heavier than TUD-Stadtmitte. Detections and ground truths are both provided by [10].

TownCenter [27]: This sequence shows a busy town center street from a single elevated camera. On average, 16 people are visible at any time, resulting in frequent dynamic occlusions. Furthermore, many people are not detected due to partial occlusions caused by static scene structures such as benches. The dataset provides manually annotated ground truth trajectories. The detection of pedestrians are coarsely estimated based on the detection of heads. We run the multi-target tracking on the first 4500 frames.

In our experiments, all parameters are fixed except $\mathbf{\Sigma}_m$. We vary it to adapt our method to different resolutions of benchmark sequences. It is reasonable in practice to tune system parameters before the deployment of a tracking algorithm according to cameras. In our experiments, we set $\mathbf{\Sigma}_m = [50, 0; 0, 50]$ for both PETS09-S2L1, PETS09-S2L2 and TUD-Stadtmitte. For both ParkingLot and TownCenter, we set $\mathbf{\Sigma}_m = [35, 0; 0, 40]$.

[1]It is worth pointing out that real-time GPU implementations exist [23]

B. Evaluation Metrics

We adopt the standard CLEAR-MOT metrics [28] for evaluation, including Multi-Object Tracking Accuracy (MOTA) and Multi-Object Tracking Precision (MOTP). We additionally report MT, ML, PT, FM, and IDS scores [29]. Finally, we provide Recall and Precision scores [30] to better evaluate different approaches. Their detailed definitions are listed below.

- **Recall**(↑): correctly matched detections / detections in the ground truth.
- **Precision**(↑): correctly matched detections / detections in the tracking results.
- **MT**(↑): the ratio of mostly tracked trajectories (tracked more than 80%). Normalized in the range $[0, 100\%]$.
- **ML**(↓): the ratio of mostly lost trajectories (tracked less than 20%). Normalized in the range $[0, 100\%]$.
- **PT**(↓): the ratio of partially tracked trajectories, *i.e.*, PT = 1 - MT - ML. Normalized in the range $[0, 100\%]$.
- **FM**(↓): the number of times the ground truth trajectory is interrupted.
- **IDS**(↓): the number of times that a trajectory changes its identity.
- **MOTA**(↑): Multi-Object Tracking Accuracy, combining missed targets, false alarms, and identity switches. Normalized in the range $[0, 100\%]$.
- **MOTP**(↑): Multi-Object Tracking Precision, the average bounding box overlap over all tracked targets. Normalized in the range $[0, 100\%]$.

↑ indicates that the higher score the better and ↓ is opposite.

C. Effectiveness of Dummy Nodes

In this paper, we propose to handle miss detections and occlusions with dummy nodes between each two consecutive frames, which does not require any occlusion reasoning. Here, we validate the effectiveness of dummy nodes by varying the threshold t_4 in Algorithm 1, which controls the maximum number of successive dummy nodes allowed in a trajectory. Beyond this threshold, a new trajectory is considered as fake and an active trajectory is considered as disappeared. By extending the standard Hungarian algorithm with the dummy node, our data association can be solved implicitly in a global manner across multiple frames although it is formulated between only two adjacent frames without considering future information. Larger t_4 allows online data associations to span more frames. In particular, when $t_4 = 0$, the dummy nodes serve only to balance the bipartite graph $\mathcal{G}$ between the frame I^t and I^{t+1} to make the standard Hungarian algorithm directly applicable. We check **IDS**, **Recall**, and **Precision** metrics on

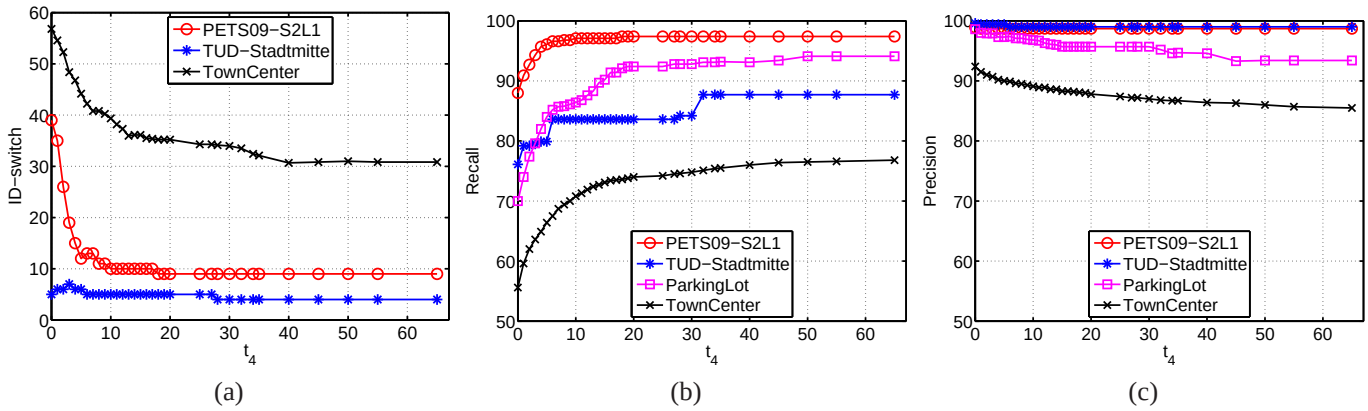

Fig. 4. Quantitative validation of the effectiveness of the dummy nodes to deal with miss detections and occlusions. From left to right: (a) IDS, (b) Recall, and (c) Precision versus the threshold t_4. The IDS on TownCenter sequence are rescaled for visualization purpose.

Fig. 5. Qualitative illustrations of the effectiveness of dummy nodes on the ParkingLot sequence with increasingly larger t_4.

the benchmark sequences, shown in Fig. 4. Since the ground truth annotations on ParkingLot sequence are given for each three frames, **IDS** are always zeros and not plotted.

As can be seen from Fig. 4(a)(b), both **IDS** and **Recall** metrics for all benchmark sequences significantly improve first and remain stable afterwards when t_4 grows. This suggests that the data association is solved implicitly in a global manner where a trajectory may find its correspondence across more frames with larger t_4. Therefore, miss detections and occlusions can be effectively handled with our proposed dummy nodes. While **Precision** remains almost steady for PETS09-S2L1 and TUD-Stadtmitte sequences as t_4 becomes larger, it slightly decreases (compared with the increase of **Recall**) for ParkingLot and TownCenter. The reason might be that some false positives are linked to trajectories when long-distance associations are allowed. Nevertheless, the decrease of **Precision** is far less than the increase of **Recall**. Therefore, by augmenting the standard Hungarian algorithm with dummy nodes, we can conclude that miss detections and occlusions can be effectively addressed during data association for multi-target tracking, where no explicit occlusion reasoning is required.

Fig. 5 also demonstrates the qualitative illustration of the effectiveness of dummy nodes to deal with miss detections and occlusions on the ParkingLot sequence. As we can observe, longer-distance data association can be addressed with the larger t_4. In specific, the association from the target highlighted with the white arrow in frame #398 is solved when $t_4 > 0$. It is also interesting to note that the very long-distance correspondence from frame #399 to frame #412 can only be found when $t_4 = 13$.

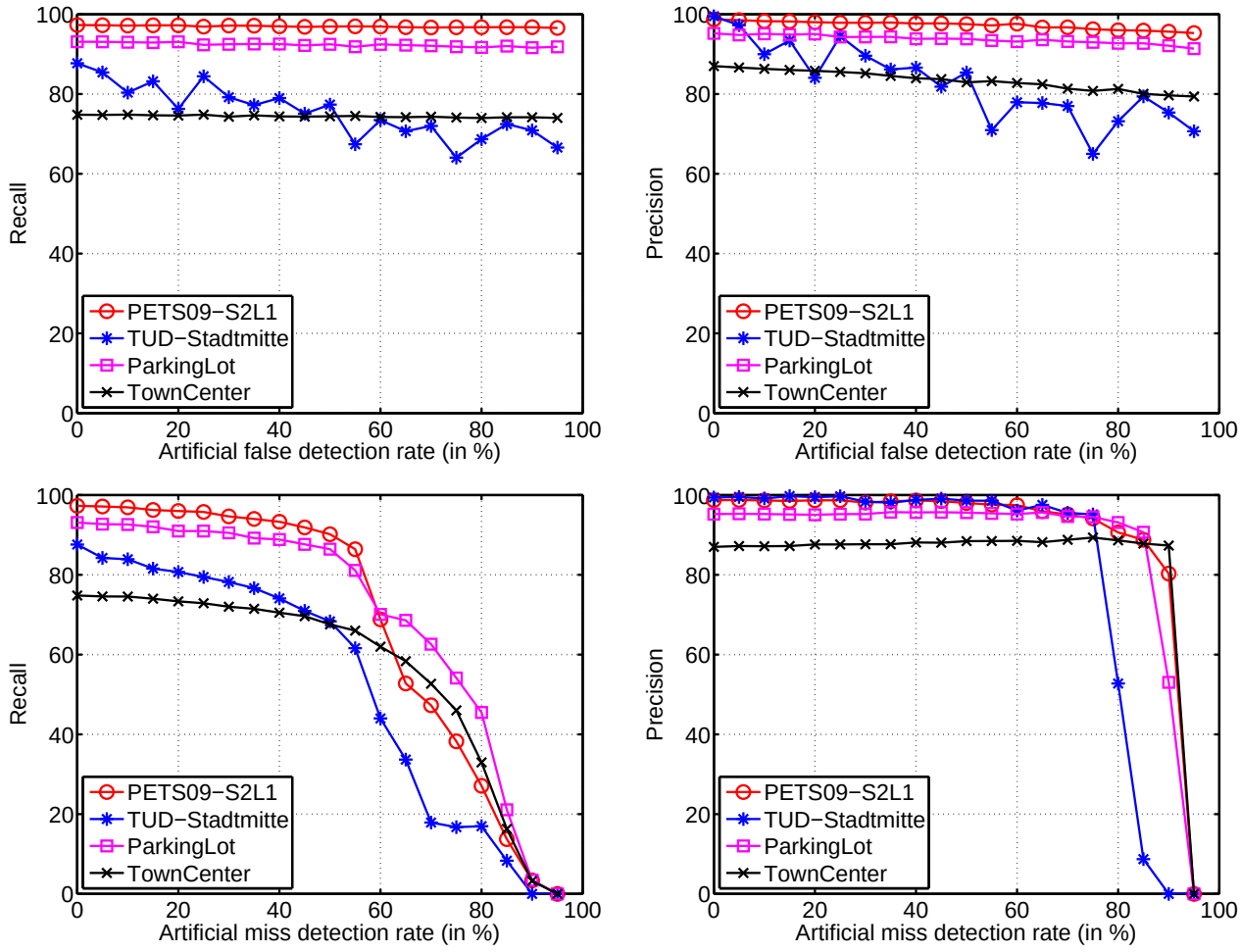

Fig. 6. Robustness analysis against **top**: false detection rates and **bottom**: miss detection rates.

D. Robustness Analysis

As we stated in Algorithm 1, there are two thresholds t_1 and t_2 to deal with false detections with dummy nodes. A set of dummy nodes will accumulate soon in a fake trajectory after its initialization since false detections are generally unstable and inconsistent. We can easily recognize a false trajectory if it is short (controlled by t_1) and has a set of successive dummy nodes in it (controlled by t_2). Instead of jointly varying them as in the previous section to check the performance variations, which might be less informative, we set them as constant and check the performance fluctuations of our proposed approach by varying the level of false detections. More importantly, robustness of our approach can be validated by doing so. Since our approach is to find correspondences of detection responses to trajectories, it can also be affected by the existence of miss detections. To this end, we quantitatively examine the robustness of our approach by generating various levels of false detection rates and miss detection rates to demonstrate robustness of our approach.

As suggested by [32], to generate controlled amounts of false detections, increasingly larger amount of random detection noises are added uniformly over the sequence. Some detections are also randomly deleted from the same original detections to generate artificially controlled miss detections. **Recall** and **Precision** metrics are reported in Fig. 6. All the experiments are conducted 10 times and the average performances are reported.

To deal with false detections, our basic assumption is that they will not appear regularly. Therefore, although a fake trajectory will be initialized with a false detection, dummy nodes will accumulate in it soon, making it easy to identify. As can be seen from the top row of Fig. 6, our approach is very robust to false detections on all benchmark sequences except TUD-Stadtmitte. Both **Recall** and **Precision** metrics decline slightly. For the TUD-Stadtmitte sequence, since the targets (pedestrians) occupy more than half of the height of the image, artificially added noises always appear in the same area and may form seemingly coherent trajectories. The performance against false detections on this sequence is thus instable. Nevertheless, we can still achieve around 0.6 in terms of both **Recall** and **Precision** measures.

As the rate of miss detections increases, it is required to consider more frames to find the correspondence of trajectories and detections, which will challenge the ability of our approach to solve data association by simply considering two consecutive frames using the dummy nodes. We can observe that our approach is robust to the miss detections under certain density value. **Recall** slightly decreases when the

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT APPROACHES ON PETS09-S2L1 SEQUENCE.

Method	Recall	Precision	MT	PT	ML	FM	IDS	MOTA	MOTP	Type
[31]	89.5	**99.6**	78.9	21.1	**0.0**	23	1	-	-	
[32]	83.8	96.3	73.9	17.4	8.7	22	13	80.3	72.0	
[6]	-	-	82.6	17.4	**0.0**	21	15	81.4	76.1	
[14]	-	-	89.5	10.5	**0.0**	45	19	83.3	71.1	
[7]	-	-	**100.0**	**0.0**	**0.0**	8	10	95.9	78.7	offline
[17]	91.8	99.0	89.5	10.5	**0.0**	9	**0**	-	-	
[33]	95.2	96.8	-	-	-	-	-	90.7	76.0	
[10]	96.5	93.6	-	-	-	-	8	90.3	69.2	
[18]			78.3	11.7	**0.0**	15	22	90.3	74.3	
[25]	92.4	98.4	91.3	4.4	4.3	6	11	90.6	80.2	
[3]	-	-	-	-	-	-	-	79.7	56.3	
[21]	-	-	**100.0**	**0.0**	**0.0**	17.5	10.5	90.1	74.3	online
[4]	-	-	**100.0**	**0.0**	**0.0**	16	9	**98.1**	80.5	
Ours	**97.8**	98.4	95.0	5.0	**0.0**	**5**	6	96.1	**87.5**	

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT APPROACHES ON PETS09-S2L2 SEQUENCE.

Method	Recall	Precision	MT	PT	ML	FM	IDS	MOTA	MOTP	Type
[12]	-	-	7.1	83.4	9.5	705	1029	33.8	**69.4**	
[18]	-	-	59.5	**21.4**	19.1	105	126	46.0	59.8	offline
[25]	**65.5**	**89.8**	37.8	46.0	16.2	**73**	**99**	56.9	59.4	
[4]	-	-	41.5	58.5	**0.0**	315	181	**66.0**	64.8	online
Ours	64.1	84.1	**59.7**	22.2	18.1	146	119	51.3	67.3	

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT APPROACHES ON TUD-STADTMITTE SEQUENCE.

Method	Recall	Precision	MT	PT	ML	FM	IDS	MOTA	MOTP	Type
[31]	81.0	**99.5**	60.0	30.0	10.0	**0**	1	-	-	
[6]	-	-	60.0	30.0	10.0	4	7	60.5	65.8	
[7]	-	-	60.0	40.0	**0.0**	1	4	61.8	63.2	offline
[9]	**87.0**	96.7	70.0	30.0	**0.0**	1	**0**	-	-	
[10]	81.4	95.6	-	-	-	-	**0**	77.7	63.4	
Ours	84.2	**99.5**	**80.0**	**20.0**	**0.0**	4	4	**83.3**	**72.2**	online

false negative rate is under around 50% while the **Precision** metric remains very steady when the miss detection rate is under around 75%. This implies that although our approach may miss some true trajectories (thus low **Recall**) due to the increasingly more miss detections, the successfully tracked trajectories are close to the ground truth (thus high **Precision**).

E. Quantitative Comparisons with State-of-the-art Methods

To verify the performance, we compare our proposed method with state-of-the-art offline and online methods. The quantitative comparisons on five benchmark sequences are presented in Table I-V, where results of other models are all based on the publicly reported literature. Since the ground truth annotations for the ParkingLot sequence are given every 3 frames, we only report the results of **Recall, Precision, MOTA**, and **MOTP**.

Comparisons with offline methods: Though our data association is formulated between each two consecutive frames, it can be implicitly solved in a global manner using dummy node as we stated in Section IV-C. Therefore comparable performances to most existing offline approaches can be achieved even if our approach only considers the past frames. For example, our method reports the best **MOTP** on the PETS09-S2L1 sequence.In particular, our approach performs worse than the state-of-the-art offline multi-target tracking methods, *e.g.*, [25] on PETS09-S2L2 and [15] on TownCenter, but slightly better than other offline methods such as [10] on both PETS09-S2L1 and Tud-Stadtmitte, and [18] on PETS09-S2L2. It is generally believed that offline tracking methods will perform better than their online counterparts because more global information (*e.g.*, high-order motion dynamics of tracklets) can be explored in the offline setting. The fact that our approach has achieved comparable performances with, but much higher running speed than most existing offline methods is a strong evidence that the proposed framework can handle the complex tracking scenarios in a unified and efficient way. It is also worth pointing out that techniques described in other papers might be complementary to ours. For example, multi-part representations [35], [33] for the pedestrians and online learned appearance [16] and motion models [17] can be further utilized to improve the performance of our method.

Comparisons with online methods: Quantitative results on the PETS09-S2L1, PETS09-S2L2, and TownCenter sequences demonstrate we can achieve slightly better performance than other online methods, *e.g.*, the best **FM** score on PETS09-S2L1. Especially on the denser PETS09-S2L2 and TownCenter sequences, occlusion reasoning might be unreliable from a single camera view [4]. With our proposed dummy nodes, miss detections and occlusions can be effectively handled. It is also worth noting that more sophisticated motion dynamics estimation is exploited in [21] to guide the correspondences of trajectories and detections.

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT APPROACHES ON PARKINGLOT SEQUENCE.

Method	Recall	Precision	MOTA	MOTP	Type
[34]	91.0	**98.5**	89.3	77.7	offline
[35]	81.7	91.3	74.1	**79.3**	
[33]	**96.5**	93.6	88.9	77.5	
[10]	94.7	93.9	88.5	77.4	
[15]	-	-	**92.9**	73.6	
Ours	93.5	95.1	88.7	75.3	online

TABLE V
QUANTITATIVE RESULTS OF DIFFERENT APPROACHES ON TOWNCENTER SEQUENCE.

Method	Recall	Precision	MT	PT	ML	FM	IDS	MOTA	MOTP	Type
[27]	79.0	82.1	60.9	27.4	11.7	**214**	293	61.3	**80.2**	offline
[35]	-	-	-	-	-	-	-	**79.3**	74.1	
[33]	**81.8**	**93.6**	-	-	-	-	-	75.7	71.6	
[10]	-	-	-	-	-	-	-	75.6	71.9	
[15]	-	-	**86.1**	**9.5**	**4.4**	-	**68**	77.4	66.4	
[21]	-	-	**64.7**	27.4	7.9	453	209	69.5	68.7	online
[4]	-	-	56.3	36.3	7.4	321	157	70.7	68.6	
Ours	74.8	87.0	50.9	33.0	16.1	356	154	63.4	72.2	

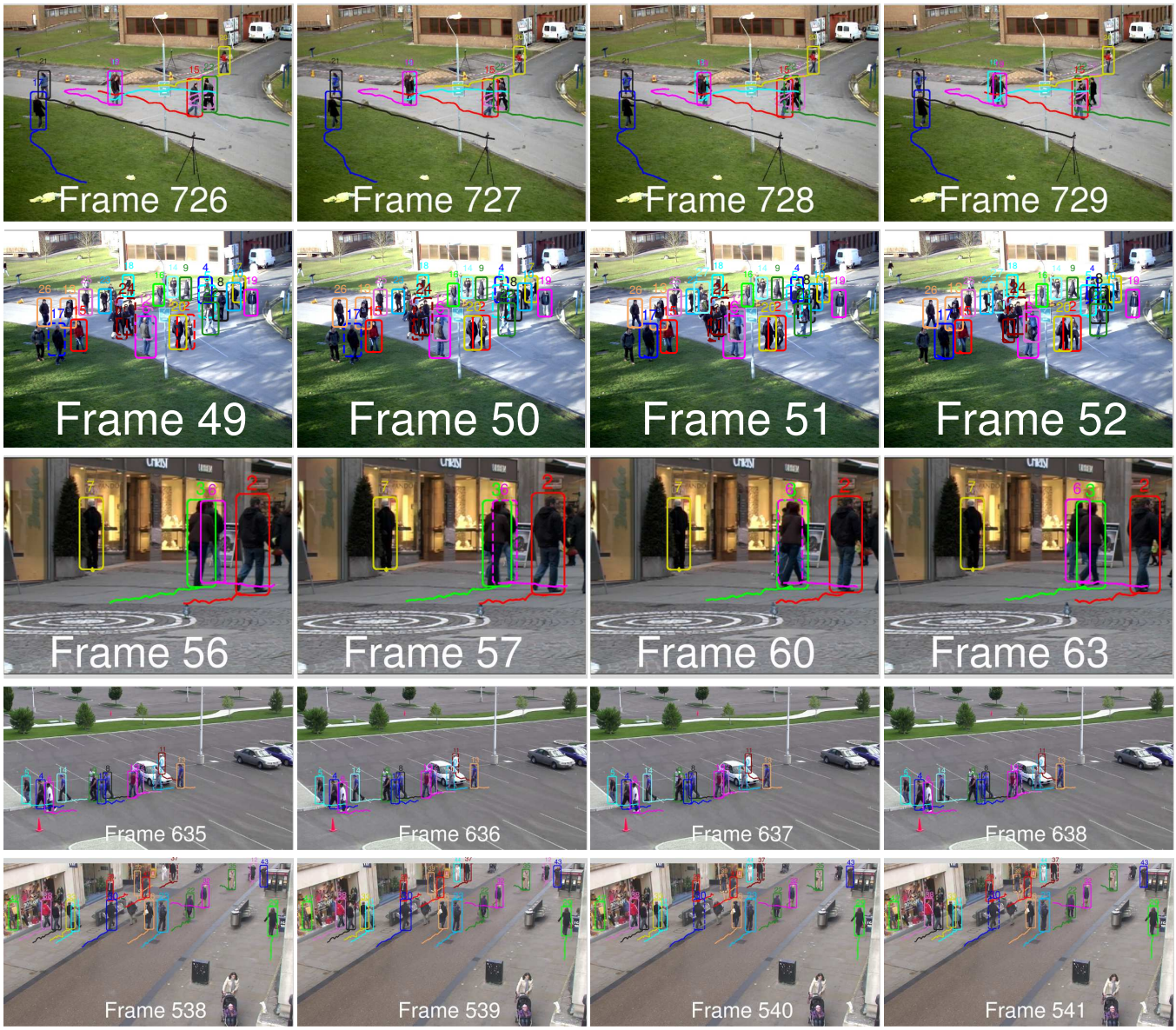

Fig. 7. Qualitative tracking results of our method on the benchmark sequences. From top to bottom: on the PETS09-S2L1, PETS09-S2L2, TUD-Stadtmitte, ParkingLot, and TownCenter, respectively.

TABLE VI
AVERAGE NUMBER OF DETECTIONS PER FRAME (IN THE SECOND COLUMN) AND PROCESSING SPEED (IN TERMS OF FPS IN THE THIRD COLUMN) OF OUR PROPOSED METHOD ON THE BENCHMARK SEQUENCES.

Dataset	#Detections	FPS
PETS09-S2L1	5.6	132
PETS09-S2L2	11.8	46
TUD-Stadtmitte	5.3	108
ParkingLot	8.8	49
TownCenter	12.0	28

F. Qulitative Results

Some qualitative tracking results produced by our proposed method on the five benchmark sequences are shown in Fig. 7.

PETS09-S2L1: Tracking results demonstrate that our proposed method can handle heavy mutual occlusions of three trajectories (trajectories #12, #15, and #22) with dummy nodes. Our method does not rely on the ambiguous occlusion reasoning (from a single uncalibrated camera view). Occlusions will be automatically handled during the association process of trajectories and detections. Occluded trajectories will be linked to their dummy nodes till the end of occlusion.

PETS09-S2L2: Our approach can successfully deal with heavy occlusions in such a crowded scene (*e.g.*, trajectory #4 in frame #50 and #52). However, some pedestrians are not tracked since they are almost fully occluded and not detected well at the beginning of the sequence.

TUD-Stadtmitte: Even if fully occluded, trajectory #6 (with magenta color) is successfully tracked. We can clearly see how our dummy nodes work under occlusions. Without explicit occlusion reasoning, our method works well using dummy nodes to handle occlusions.

ParkingLot: Pedestrians gradually enter the camera view and finally move out. Our method is able to handle the miss detection of trajectory #10 in frame #638 (with canyon color). Occlusion of trajectory #11 (with dark read) by the car is successfully handled as well.

TownCenter: In this very crowded sequence, pedestrians frequently enter and leave the camera view. Miss detections and occlusions are successfully dealt with using dummy nodes. Additionally, new trajectories (*e.g.*, trajectory #44) can be automatically discovered though we do not utilize scene priors.

Trajectory initializations and terminations happen anytime and anywhere in each benchmark sequence. Without the assistance of scene prior knowledge, our approach succeeds in automatically discovering new targets and terminating disappeared trajectories.

G. Runtime Performance

A computationally cheap approach is always preferred to track multiple targets, especially when dealing with a large amount of video frames. Our single-thread code is written in C++ without any particular optimization and tested on a desktop equipped by an Intel i7 CPU with 3.4GHz and 32GB of memory. Runtime performance of our proposed approach on benchmark sequences are summarized in Table VI.

Our approach does not require the time-consuming occlusion reasoning. Additionally, we only need to check the number of consecutive dummy nodes to handle false detections and trajectory terminations. Therefore, our approach is computationally cheap, running at 132 fps (frame per second) on a typical scene with moderate crowdedness (*e.g.*, PETS09-S2L1) excluding the target detection step. In contrast, state-of-the-art multi-target trackers report substantially slower processing speeds[2], 0.4-2 fps [3] (implemented with C++), 1-2 fps [21] (implemented with MATLAB), and 11.2 fps [4] (implemented with MATLAB), also without considering the detection step. Even on the crowded PETS09-S2L2 and TownCenter sequences, our proposed method runs at 46 and 28 fps, respectively. This suggests that in a whole tracking-by-detection system, our proposed multi-target tracking method will not be the bottleneck of the processing speed and suitable for time-critical scenarios (*e.g.*, video surveillance).

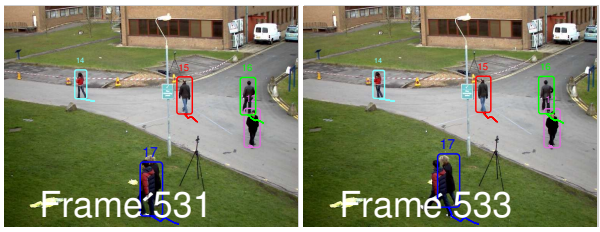

Fig. 8. A failure case of our method which generates an ID switch between two trajectories.

V. CONCLUSION

In this paper, we propose an online tracking-by-detection method for multi-target tracking from a single uncalibrated camera. With proposed dummy nodes, complex scenarios including miss detections, occlusions, false detections, and trajectory terminations can be handled in a unified manner. Though data association is formulated between each two consecutive frames, it is implicitly solved in a global manner. Moreover, our proposed method does not require explicit occlusion reasoning, which might be time consuming and ambiguous from a single camera view, leading to an efficient multi-target tracking method. It runs at 132 fps on the PETS09-S2L1 benchmark sequence. Last but not least, our proposed method does not necessarily rely on scene priors, *e.g.*, the entrance and exit area of trajectories. Thus it has more wider applications. Quantitative comparisons on five benchmark sequences demonstrate that we can achieve comparable results with most existing offline methods and better results than other online algorithms.

In Fig. 8, we present a failure case of our method. The association of the trajectory #17 in frame #532 fails to find its matching detection and finally drifts in frame #533. An ID switch is generated. This is because the motion cue in this case is not reliable and the appearance cue (color histogram) is not discriminative enough though these two targets look very distinct. This motivates us to adopt more powerful appearance cues, *e.g.*, the online learned appearance model [16]. We leave this as our future work.

[2]Detailed comparisons with state-of-the-art methods are not feasible, however, since their codes are not available. Running speed of most of the approaches are not explicitly reported in the literature either.

We found that a Multiple Object Tracking Benchmark [36] (http://motchallenge.net/) was established recently. As a future work, we plan to run evaluations on the entire 2D tracking benchmark sequences.

ACKNOWLEDGMENT

The first author would like to thank Shun Zhang for proofreading the draft. This work was supported by the National Basic Research Program of China under Grant No. 2015CB351705, and the National Natural Science Foundation of China under Grant No. 61332018.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR (1)*, 2005, pp. 886–893.
[2] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.
[3] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. J. V. Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1833, 2011.
[4] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof, "Occlusion geodesics for online multi-object tracking," in *CVPR*, 2014.
[5] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *CVPR*, 2008.
[6] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *CVPR*, 2011, pp. 1265–1272.
[7] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *CVPR*, 2012, pp. 1926–1933.
[8] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *ECCV*, 2008, pp. 788–801.
[9] B. Yang and R. Nevatia, "An online learned crf model for multi-target tracking," in *CVPR*, 2012, pp. 2034–2041.
[10] A. R. Zamir, A. Dehghan, and M. Shah, "Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs," in *ECCV*, 2012, pp. 343–356.
[11] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
[12] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *CVPR*, 2011, pp. 1201–1208.
[13] A. A. Butt and R. T. Collins, "Multi-target tracking by lagrangian relaxation to min-cost network flow," in *CVPR*, 2013, pp. 1846–1853.
[14] J. F. Henriques, R. Caseiro, and J. Batista, "Globally optimal solution to multi-object tracking with merged measurements," in *ICCV*, 2011, pp. 2470–2477.
[15] A. Dehghan, S. M. Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *CVPR*, 2015.
[16] C.-H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *CVPR*, 2010, pp. 685–692.
[17] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *CVPR*, 2012, pp. 1918–1925.
[18] A. Milan, K. Schindler, and S. Roth, "Detection- and trajectory-level exclusion in multiple object tracking," in *CVPR*, 2013, pp. 3682–3689.
[19] S. Pellegrini, A. Ess, K. Schindler, and L. J. V. Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *ICCV*, 2009, pp. 261–268.
[20] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *CVPR*, 2011, pp. 1345–1352.
[21] Z. Wu, J. Zhang, and M. Betke, "Online motion agreement tracking," in *BMVC*, 2013, pp. 1–10.
[22] G. Welch and G. Bishop, "An introduction to the kalman filter," Chapel Hill, NC, USA, Tech. Rep., 1995.
[23] V. Prisacariu and I. Reid, "fasthog - a real-time gpu implementation of hog," Department of Engineering Science, Oxford University, Tech. Rep. 2310/09.
[24] A. Ellis and J. M. Ferryman, "Pets2010 and pets2009 evaluation of results using individual ground truthed single views," in *AVSS*, 2010, pp. 135–142.
[25] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 58–72, 2014.
[26] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *CVPR*, 2008.
[27] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR*, 2011, pp. 3457–3464.
[28] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The clear 2006 evaluation," in *Multimodal Technologies for Perception of Humans*. Springer, 2007, pp. 1–44.
[29] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," in *CVPR*, 2006, pp. 951–958.
[30] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *CVPR*, 2009, pp. 2953–2960.
[31] C.-H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?" in *CVPR*, 2011, pp. 1217–1224.
[32] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, 2011.
[33] H. Izadinia, I. Saleemi, W. Li, and M. Shah, "(mp)2t: Multiple people multiple parts tracker," in *ECCV (6)*, 2012, pp. 100–114.
[34] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele, "Learning people detection models from few training samples," in *CVPR*, 2011, pp. 1473–1480.
[35] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *CVPR*, 2012, pp. 1815–1821.
[36] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015, arXiv: 1504.01942. [Online]. Available: http://arxiv.org/abs/1504.01942

Huaizu Jiang is a research assistant at Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. Before that, he received his BS and MS degrees from Xi'an Jiaotong University, China, in 2005 and 2009, respectively. He is interested in how to teach an intelligent machine to understand the visual scene like a human. Specifically, his research interests include object detection, large-scale visual recognition, and (3D) scene understanding.

Jinjun Wang is a professor with Xian Jiaotong university. He received the B.E. and M.E. degrees from Huazhong University of Science and Technology, China, in 2000 and 2003, and received the Ph.D degree from Nanyang Technological University, Singapore, in 2006. From 2006 to 2009, Dr. Wang was with NEC Laboratories America, Inc. as a research scientist, and from 2010 to 2013, he was with Epson Research and Development, Inc. as a senior research scientist. Dr. Wangs research interests include visual classification, content-based image/video annotation and retrieval, image/video enhancement and editing, etc.

Yihong Gong received his B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Tokyo, Japan in 1987, 1989, and 1992, respectively. In 1992, he joined Nanyang Technological University, Singapore as an assistant professor with the School of Electrical and Electronic Engineering. From 1996 to 1998, he was a Project Scientist with the Robotics Institute, Carnegie Mellon University, USA. Since 1999 he worked for the Silicon Valley branch, NEC Labs America as a group leader, department head, and branch manager. In 2012, he joined Xian Jiaotong University, China as a distinguished professor. His research interests include image and video analysis, multimedia database systems, and machine learning.

Na Rong received her B.S. and M.S. degrees from Xi'an Jiaotong University, China, in 2013 and 2015, respectively. She is interested in object tracking and person re-identification.

Zhenhua Chai is currently a Senior Research Engineer in Media Technology Lab, Central Research Institute at Huawei Technologies Co.,Ltd. He received the B.E. degree in Automation (with Honors) from Central University of Nationality (CUN) and the Ph.D. degree in Computer Application Technology from National Lab of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) in 2008 and 2013 respectively. His research interests focus on face recognition, feature extraction and system integration. He has published several papers on feature extraction for face recognition.

Nanning Zheng (SM93-F06) graduated from the Department of Electrical Engineering, Xian Jiaotong University, Xian, China, in 1975, and received the M.S. degree in information and control engineering from Xian Jiaotong University in 1981 and the Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1985. He jointed Xian Jiaotong University in 1975, and he is currently a Professor and the Director of the Institute of Artificial Intelligence and Robotics, Xian Jiaotong University. His research interests include computer vision, pattern recognition and image processing, and hardware implementation of intelligent systems. Dr. Zheng became a member of the Chinese Academy of Engineering in 1999, and he is the Chinese Representative on the Governing Board of the International Association for Pattern Recognition. He also serves as an executive deputy editor of the Chinese Science Bulletin.