Robust Multi-Object Tracking by Marginal Inference

Yifu Zhang[1†], Chunyu Wang[2], Xinggang Wang[1], Wenjun Zeng[3], and Wenyu Liu[1‡]

[1] Huazhong University of Science and Technology
[2] Microsoft Research Asia
[3] Eastern Institute for Advanced Study

Abstract. Multi-object tracking in videos requires to solve a fundamental problem of one-to-one assignment between objects in adjacent frames. Most methods address the problem by first discarding impossible pairs whose feature distances are larger than a threshold, followed by linking objects using Hungarian algorithm to minimize the overall distance. However, we find that the distribution of the distances computed from Re-ID features may vary significantly for different videos. So there isn't a single optimal threshold which allows us to safely discard impossible pairs. To address the problem, we present an efficient approach to compute a marginal probability for each pair of objects in real time. The marginal probability can be regarded as a normalized distance which is significantly more stable than the original feature distance. As a result, we can use a single threshold for all videos. The approach is general and can be applied to the existing trackers to obtain about one point improvement in terms of IDF1 metric. It achieves competitive results on MOT17 and MOT20 benchmarks. In addition, the computed probability is more interpretable which facilitates subsequent post-processing operations.

Keywords: multi-object tracking, data association, marginal probability

1 Introduction

Multi-object tracking (MOT) is one of the most active topics in computer vision. The state-of-the-art methods [48,45,58,60,25,31,37,40] usually address the problem by first detecting objects in each frame, and then linking them to trajectories based on Re-ID features. Specifically, it computes distances between objects in adjacent frames, discards impossible pairs with large distances, and determines the matched pairs by minimizing the overall distance by applying the Hungarian algorithm [20].

The core of the linking step is to find a threshold where the distances between the matched objects are smaller than it, while those of the unmatched ones are larger than it. The threshold setting is done by experience and has not received sufficient attention. However, our experiment shows that even the best Re-ID model cannot discard all impossible pairs without introducing false negatives using a single threshold because the distances may vary significantly on different frames as shown in Figure 1 (top). We can see that the optimal threshold to discriminate matched and unmatched pairs is 0.2 for video "MOT17-04" and 0.4 for "MOT17-09" which are very different.

† This work was done when Yifu Zhang was an intern of Microsoft Research Asia. ‡ Corresponding author.

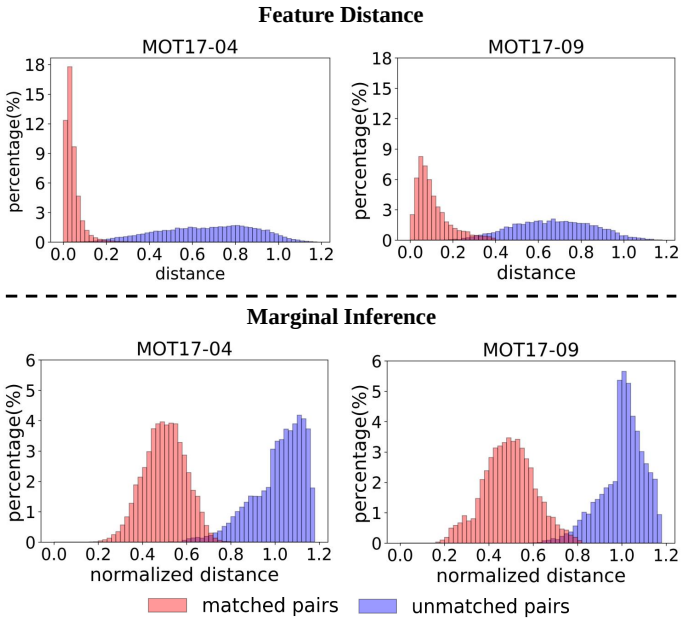

Fig. 1. Distance distribution of the matched pairs and unmatched pairs, respectively, on two videos. The top shows the distances directly computed from Re-ID features the bottom shows our normalized distances (marginal probability).

We argue in this work that we should put a particular value of distance into context when we determine whether it is sufficiently small to be a matched pair. For example, in Figure 1 (top), 0.2 is a large distance for video "MOT17-04" but it is a small one for "MOT17-09" considering their particular distance distributions. To achieve this goal, we propose to compute a marginal probability for each pair of objects being matched by considering the whole data association space which consists of all possible one-to-one assignment structures. The marginal probability is robust to distance distribution shift and significantly improves the linking accuracy in our experiment. For example, in the "Public Detection" track of the MOT17 challenge, the IDF1 score improves from 59.6% to 65.0%.

We consider a possible matching between all the detections and trajectories as one structure. However, naively enumerating all structures is intractable especially when the number of objects in videos is large. We address the complexity issue by computing a small number of low-cost supporting structures which often overlap with the maximum a posterior solution found by Hungarian algorithm [4,48]. The marginal probability of each pair is computed by performing marginal inference among these structures. Our experiments on videos with a large number of objects show that it takes only a fraction of the time to compute without affecting the inference speed.

The approach is general and applies to almost all existing multi-object trackers. We extensively evaluate approach with the state-of-the-art trackers on multiple datasets. It consistently improves the tracking performances of all methods on all datasets with little

extra computation. In particular, we empirically find that it is more robust to occlusion. When occlusion occurs, the distance between two (occluded) instances of the same person becomes larger and the conventional methods may treat them as two different persons. However, the marginal probability is less affected and the two instances can be correctly linked.

2 Related work

Most *state-of-the-art* multi-object tracking methods [48,45,58,60,25,31,37,56,51] follow the *tracking-by-detection* paradigm which form trajectories by associating detections in time. They first adopt detectors such as [34,33,61,22,12] to get the location of the objects and then link the detections to the existing trajectories according to similarity. Similarity computation and matching strategy are two key components of data association in the *tracking-by-detection* paradigm. We review different methods from the two aspects and compare them to our approach.

2.1 Similarity computation

Location, motion, and appearance are three important cues to compute similarity between detections and tracks. IOU-Tracker [6] computes the spatial overlap of detections in neighboring frames as similarity. SORT [4] adopts Kalman Filter [47] as a motion model to predict the future locations of objects. The similarity is computed by the IoU of predicted locations and detected locations. The two trackers are widely used in practice due to their speed and simplicity. However, both the two trackers will cause a large number of identity switches when encountering camera motion and crowded scenes. To decrease the identity switches, DeepSORT [48] adopts a deep neural network to extract appearance features as appearance cues can refind lost objects. The final similarity is a weighted sum of motion similarity computed by Kalman Filter and cosine similarity of the appearance features. Bae *et al*. [1] also proposes an online discriminative appearance learning method to handle similar appearances of different objects. Many *state-of-the-art* methods [9,42,45,58,25,31,57] follow [48] to compute the similarity using location, motion and appearance cues. Some methods [49,37] utilize networks to encode appearance and location cues into similarity score. Recently, some methods [45,58,25,31] combine the detection task and re-identification task in a single neural network to reduce computation cost. JDE [45] first proposes a joint detection and embedding model to first achieve a (near) real-time MOT system. FairMOT [58] deeply studies the reasons for the unfairness between detection and re-identification task in anchor-based models and proposes a high-resolution anchor-free model to extract more discriminative appearance features. QDTrack [31] densely samples hundreds of region proposals on a pair of images for contrastive learning of appearance features to make use of the majority of the informative regions on the images. It gets very high-quality appearance features and can achieve *state-of-the-art* results only using appearance cues.

Our method also uses the location, motion, and appearance cues to compute similarity. However, we find that the appearance feature distance distribution may vary significantly for different videos. We achieve a more stable distribution by computing the marginal probability based on appearance features.

2.2 Matching strategy

After computing similarity, most methods [4,48,1,9,42,45,58,49,37] use Hungarian Algorithm [20] to complete matching. Bae *et al*. [1] matches tracklets in different ways according to their confidence values. Confident tracklets are locally matched with online-provided detections and fragmented tracklets are globally matched with confident tracklets or unmatched detections. The advantage of confidence-based tracklets matching is that it can handle track fragments due to occlusion or unreliable detections. DeepSORT [48] proposes a cascade matching strategy, which first matches the most recent tracklets to the detections and then matches the lost tracklets. This is because recent tracklets are more reliable than lost tracklets. MOTDT [9] proposes a hierarchical matching strategy. It first associates using the appearance and motion cues. For the unmatched tracklets and detections (usually under severe occlusion), it matches again by IoU. JDE [45] and FairMOT [58] also follow the hierarchical matching strategy proposed by MOTDT. QDTrack [31] applies a bi-directional softmax to the appearance feature similarity and associates objects with a simple nearest neighbor search. All these methods need to set a threshold to decide whether the detections match the tracklets. If the distance is larger than the threshold, the matching is rejected. It is very challenging to set an optimal threshold for all videos because the distance distribution may vary significantly as the appearance model is data-driven. The work most related to our approach is Rezatofighi *et al*. [35] which also uses probability for matching. However, their motivation and the solution to compute probability are different from ours.

We follow the matching strategy of [9,45,58] which hierarchically uses appearance, motion, and location cues. The main difference is that we turn the appearance similarity into marginal probability (normalized distance) to achieve a more stable distance distribution. The motion model and location cues are more generalized, so we do not turn them into probability. The matching is also completed by the Hungarian Algorithm. Our marginal probability is also more robust to occlusion as it decreases the probability of false matching. We can thus set a looser (higher) matching threshold to refind some lost/occluded objects.

3 Method

3.1 Problem formulation

Suppose we have M detections and N history tracks at frame t, our goal is to assign each detection to one of the tracks which has the same identity. Let $\mathbf{d}_t^1, ..., \mathbf{d}_t^M$ and $\mathbf{h}_t^1, ..., \mathbf{h}_t^N$ be the Re-ID features of all M detections and N tracks at frame t, respectively. We compute a cosine similarity matrix $\mathbf{S}_t \in [0,1]^{M \times N}$ between all the detections and tracks as follows:

$$\mathbf{S}_t(i,j) = \frac{\mathbf{d}_t^i \cdot \mathbf{h}_t^j}{||\mathbf{d}_t^i|| \cdot ||\mathbf{h}_t^j||}, \tag{1}$$

where $i \in \{1, ..., M\}$ and $j \in \{1, ..., N\}$. For simplicity, we replace $\{1, ..., M\}$ by $\mathbb{M}$ and $\{1, ..., N\}$ by $\mathbb{N}$ in the following myparagraphs.

Based on the similarity $\mathbf{S}_t$, we compute a marginal probability matrix $\mathbf{P}_t \in [0,1]^{M \times N}$ for all pairs of detections and tracks. $\mathbf{P}_t(i,j)$ represents the marginal probability that the

i_{th} detection is matched to the j_{th} track. We compute $\mathbf{P}_t(i,j)$ considering all possible matchings. Let $\mathbb{A}$ denote the space which consists of all possible associations (or matchings). Under the setting of multi-object tracking, each detection matches at most one track and each track matches at most one detection. We define the space $\mathbb{A}$ as follow:

$$\mathbb{A} = \Big\{ A = \Big(m_{ij}\Big)_{i\in\mathbb{M}, j\in\mathbb{N}} \;\Big|\; m_{ij} \in \{0,1\} \tag{2}$$

$$\wedge \sum_{i=0}^{M} m_{ij} \leqslant 1, \forall j \in \mathbb{N} \tag{3}$$

$$\wedge \sum_{j=0}^{N} m_{ij} \leqslant 1, \forall i \in \mathbb{M} \Big\}, \tag{4}$$

where A is one possible matching. We define $\mathbb{A}_{ij}$ as a subset of $\mathbb{A}$, which contains all the matchings where the i_{th} detection is matched to the j_{th} track:

$$\mathbb{A}_{ij} = \{A \in \mathbb{A} \mid m_{ij} = 1\} \tag{5}$$

The marginal probability $\mathbf{P}_t(i,j)$ can be computed by marginalizing $\mathbb{A}_{ij}$ as follows:

$$\mathbf{P}_t(i,j) = \sum_{A\in\mathbb{A}_{ij}} p(A), \tag{6}$$

where $p(A)$ is a joint probability representing the probability of one possible matching A and can be computed as follows:

$$p(A) = \prod_{\forall q\in\mathbb{M}, \forall r\in\mathbb{N}} \left(\frac{\exp\big(\mathbf{S}_t(q,r)\big)}{\sum_{r=1}^{N} \exp\big(\mathbf{S}_t(q,r)\big)}) \right) \tag{7}$$

The most difficult part to obtain $\mathbf{P}_t(i,j)$ is to computing all possible matchings in $\mathbb{A}_{ij}$ because the total number of matchings is *n*-permutation.

3.2 Our solution

We consider the structured problems. We view each possible matching between all the detections and tracks as a structure $\mathbf{k} \in \{0,1\}^{MN}$, which can be seen as a flattening of the matching matrix. We define all the structures in the space $\mathbb{A}$ as $\mathbf{K} \in \{0,1\}^{MN\times D}$, where D is the number of all possible matchings and $MN \ll D$.

We often use the structured log-potentials to parametrize the structured problems. The scores of the structures can be computed as $\boldsymbol{\theta} := \mathbf{K}^\top \boldsymbol{S}$, where $\boldsymbol{S} \in \mathbb{R}^{MN}$ is a flattening of the similarity matrix $\mathbf{S}_t$. Suppose we have variables V and factors F in a factor graph [19], $\boldsymbol{\theta}$ can be computed as:

$$\theta_o := \sum_{v\in V} S_{V,v}(o_v) + \sum_{f\in F} S_{F,f}(o_f), \tag{8}$$

where o_v and o_f are local structures at variable and factor nodes. $\boldsymbol{S}_V$ and $\boldsymbol{S}_F$ represent the log-potentials. In our linear assignment setting, we only have variables and thus $\boldsymbol{\theta}$ can be written in matrix notation as $\boldsymbol{\theta} = \mathbf{K}^\top \boldsymbol{S}_V$.

The optimal matching between detections and tracks can be viewed as the MAP inference problem, which seeks the highest-scoring structure. It can be rewritten using the structured log-potentials as follows:

$$\text{MAP}_{\mathbf{K}}(\boldsymbol{S}) := \underset{\mathbf{v}:=\mathbf{K}\mathbf{y}, \mathbf{y}\in\triangle^D}{\arg\max}\ \boldsymbol{\theta}^\top \mathbf{y} \tag{9}$$

$$= \underset{\mathbf{v}:=\mathbf{K}\mathbf{y}, \mathbf{y}\in\triangle^D}{\arg\max}\ \boldsymbol{S}_V{}^\top \mathbf{v}, \tag{10}$$

where $\mathbf{v} \in \{0,1\}^{MN}$ is the highest-scoring structure and $\{\mathbf{v} = \mathbf{K}\mathbf{y}, \mathbf{y} \in \triangle^D\}$ is the Birkhoff polytope [5]. In linear assignment, the structure $\mathbf{v}$ can be obtained by Hungarian algorithm [20].

The main challenge of computing the marginal probability as in Equation 6 is that the total number of structures D is very large and usually not tractable. To address the problem, we propose to compute a small number of low-cost and often-overlapping structures instead of enumerating all of them. In [29], Niculae *et al*. show that this can be achieved by regularizing the MAP inference problem with a squared l_2 penalty on the returned posteriors which was inspired by [27]. Computing multiple structures which approximate the MAP inference problem can be written as follows:

$$\text{L2MAP}_{\mathbf{K}}(\boldsymbol{S}) := \underset{\mathbf{v}:=\mathbf{K}\mathbf{y}, \mathbf{y}\in\triangle^D}{\arg\max}\ \boldsymbol{\theta}^\top \mathbf{y} - \frac{1}{2}||\mathbf{K}\mathbf{y}||_2^2 \tag{11}$$

$$= \underset{\mathbf{v}:=\mathbf{K}\mathbf{y}, \mathbf{y}\in\triangle^D}{\arg\max}\ \boldsymbol{S}_V{}^\top \mathbf{v} - \frac{1}{2}||\mathbf{v}||_2^2, \tag{12}$$

The result is a quadratic optimization problem and it can be solved by the conditional gradient (CG) algorithm [21]. The Equation 11 can be written by function f as follows:

$$f(\mathbf{v}) := \boldsymbol{S}_V{}^\top \mathbf{v} - \frac{1}{2}||\mathbf{v}||_2^2, \tag{13}$$

A linear approximation to f around a point $\mathbf{v}'$ is:

$$\hat{f}(\mathbf{v}) := (\nabla_{\mathbf{v}} f)^\top \mathbf{v} = (\boldsymbol{S}_V - \mathbf{v}')^\top \mathbf{v}, \tag{14}$$

We can turn the optimization problem of $\hat{f}$ into an MAP inference problem. The variable scores of the MAP inference problem at each step is $\boldsymbol{S}_V - \mathbf{v}'$. At each step, we use Hungarian algorithm to solve the MAP inference problem and get a high-scoring structure $\mathbf{z} \in \{0,1\}^{MN}$. Then we use $\mathbf{z}$ to substitute $\mathbf{v}'$ for another step. After a small number of steps, we obtain a set of high-scoring and often-overlapping structures $\mathbb{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$, where n is the number of steps. We compute the marginal probability $\mathbf{P}_t$ by marginalizing $\mathbb{Z}$ following Equation 6:

$$\mathbf{P}_t(i,j) = \sum_{\mathbf{z}\in\mathbb{Z}_{ij}} \left(\frac{\exp(-\mathbf{C}_t^\top \mathbf{z})}{\sum_{v=1}^{n} \exp(-\mathbf{C}_t^\top \mathbf{z}_v)} \right) \tag{15}$$

where $\mathbb{Z}_{ij}$ contains all structures that the i_{th} detection is matched to the j_{th} track and $\mathbf{C}_t \in [0,1]^{MN}$ is a flattened feature distance matrix $\boldsymbol{S}_V$.

3.3 Tracking algorithm

Our tracking algorithm jointly considers appearance, motion, and location cues. In each frame, we adopt Hungarian algorithm [20] to perform matching two times hierarchically: 1) marginal probability matching, 2) IoU matching.

We first define some thresholds in our tracking algorithm. Thresholds C_t and C_d are the confidence thresholds for the detections. Thresholds T_p and T_{IoU} are for the marginal probability matching and IoU matching, respectively.

In the first frame, we initialize all detections with scores larger than C_d as new tracks. In the following frame t, we first match between all N tracks and all M medium detections using marginal probability $\mathbf{P}_t \in [0,1]^{M\times N}$ calculated by Equation 15 and the Mahalanobis distance $\mathbf{M}_t \in \mathbb{R}^{M\times N}$ computed by Kalman Filter proposed in [48]. The cost matrix is computed as follows:

$$\mathbf{D}_p = \omega(1-\mathbf{P}_t) + (1-\omega)\mathbf{M}_t, \tag{16}$$

where ω is a weight that balances the appearance cue and the motion cue. We set ω to be 0.98 in our experiments. We adopt Hungarian algorithm to perform the first matching and we reject the matching whose distance is larger than T_p. It is worth noting that T_p may vary significantly for different frames or videos if we directly utilize the appearance feature similarity for matching. The marginal probability matching has a significantly more stable T_p.

For the unmatched detections and tracks, we perform the second matching using IoU distance with the threshold T_{IoU}. This works when appearance features are not reliable (*e.g.* occlusion).

Finally, we mark lost for the unmatched tracks and save it for 30 frames. For the unmatched detections, if the score is larger than C_t, we initialize a new track. We also update the appearance features following [58].

4 Experiments

4.1 MOT benchmarks and metrics

Datasets. We evaluate our approach on MOT17 [28] and MOT20 [11] benchmarks. The two datasets both provide a training set and a test set, respectively. The MOT17 dataset has videos captured by both moving and stationary cameras from various viewpoints at different frame rates. The videos in the MOT20 dataset are captured in very crowded scenes so there is a lot occlusion happening. There are two evaluation protocols which either use the provided public detections or private detections generated by any detectors. In particular, the MOT17 dataset provides three sets of public detections generated DPM [13], Faster R-CNN [34] and SDP [52], respectively and we evaluate our approach on all of them. The MOT20 dataset provides one set of public detections generated by Faster R-CNN.

Metrics. We use the CLEAR metric [3] and IDF1 [36] to evaluate different aspects of multi-object tracking. Multi-Object Tracking Accuracy (MOTA) and Identity F1 Score (IDF1) are two main metrics. MOTA focuses more on the detection performance. IDF1 focuses on identity preservation and depends more on the tracking performance.

4.2 Implementation details

We evaluate our approach with two existing feature extractors. The first is the state-of-the-art one-stage method FairMOT [58] which jointly detects objects and estimates Re-ID features in a single network. The second is the state-of-the-art two-stage method following the framework of DeepSORT [48] which adopts Scaled-YOLOv4 [43] as the detection model and BoT [26] as the Re-ID model. The input image is resized to 1088×608. In the linking step, we set $C_d = 0.4$, $C_t = 0.5$, $T_p = 0.8$, $T_{IoU} = 0.5$. We set the number of steps n to be 100 when computing the marginal probability. The inference speed of the models is listed as follows: 40 FPS for Scaled-YOLOv4, 26 FPS for FairMOT and 17 FPS for BoT.

Public detection. In this setting, we adopt FairMOT [58] and our marginal inference data association method. Following the previous works of Tracktor [2] and CenterTrack [60], we only initialize a new trajectory if it is near a public detection (*e.g.* IoU is larger than a threshold). In particular, we set a strict IoU theshold 0.75 to make our trajectories as close to the public detections as possible. The FairMOT model is pre-trained on the COCO dataset [23] and finetuned on the training set of MOT17 and MOT20, respectively.

Private detection. We adopt the detection model Scaled-YOLOv4 [43] and Re-ID model BoT [26] implemented by FastReID [14] in the private detection setting. We train Scaled-YOLOv4 using the YOLOv4-P5 [43] model on the same combination of different datasets as in FairMOT [58]. We train BoT on Market1501 [59], DukeMTMC [36] and MSMT17 [46]. All the training process is the same as the references except the training data. For the Re-ID part, we multiply the cosine distance by 500 to make it evenly distributed between 0 and 1.

Ablation study. For ablation study, we evaluate on the training set of MOT17 and MOT20. To make it more similar to real-world applications, our training data and evaluation data have different data distribution. We also use different detection models and Re-ID models to evaluate the generalization ability of our method. We select three models, FairMOT, Scaled-YOLOv4, and BoT. We adopt FairMOT either as a joint detection and Re-ID model or a separate Re-ID model. We adopt Scaled-YOLOv4 as a detection model and BoT as a Re-ID model. We train Scaled-YOLOv4 on the CrowdHuman [38] dataset. We train FairMOT on the HiEve [24] dataset. We train BoT on the Market1501, DukeMTMC, and MSMT17 datasets. The matching threshold is 0.4 for the distance-based method and 0.8 for the probability-based method.

4.3 Evaluation of the marginal probability

More stable distribution. In this part, we try to prove that marginal probability is more stable than feature distance. We compare the IDF1 score of the two different methods and also plot the distance distribution between detections and tracks of each method. We adopt different detection models and Re-ID models and evaluate on different datasets. The results are shown in Table 1. We can see that the marginal probability matching method has about 1 point IDF1 score higher than the distance-based matching method

in most settings. The optimal threshold for each video is different. However, it is not realistic to set different thresholds for different videos in some testing scenarios or real-world applications. So we set one threshold for all videos in a dataset. One reason for the performance gain is that the optimal threshold for each video is similar in the probability-based method, which means a single threshold is suitable for most videos.

To obtain the distance distribution, we adopt Scaled-YOLOv4 [43] as the detector and FairMOT [58] as the Re-ID model to perform tracking on the training set of MOT17. We find the optimal threshold for each video by grid search and plot the distance distribution between detections and tracks in each video. The saved distance is the input of the first matching in our tracking algorithm. As is shown in Figure 1, the marginal probability distribution of each video is similar and is more stable than feature distance distribution. Also, the optimal threshold for each video is similar in the probability-based matching method and varies significantly in the distance-based matching method.

Dataset	Det	ReID	Match	MOTA↑	IDF1↑
MOT17	FairMOT	FairMOT	D	47.6	58.0
MOT17	FairMOT	FairMOT	P	47.6	**58.2 (+0.2)**
MOT17	Scaled-YOLOv4	FairMOT	D	71.5	72.5
MOT17	Scaled-YOLOv4	FairMOT	P	71.4	**73.6 (+1.1)**
MOT17	Scaled-YOLOv4	BoT	D	71.5	74.8
MOT17	Scaled-YOLOv4	BoT	P	71.6	**75.7 (+0.9)**
MOT20	FairMOT	FairMOT	D	43.2	45.9
MOT20	FairMOT	FairMOT	P	43.1	**46.7 (+0.8)**
MOT20	Scaled-YOLOv4	FairMOT	D	73.9	69.3
MOT20	Scaled-YOLOv4	FairMOT	P	74.1	**70.1 (+0.8)**
MOT20	Scaled-YOLOv4	BoT	D	74.0	69.5
MOT20	Scaled-YOLOv4	BoT	P	74.2	**70.2 (+0.7)**

Table 1. Comparison of probability-based method and distance-based method. "P" is short for probability and "D" is short for distance. "Det" is short for detection model and "Re-ID" is short for Re-ID model.

More robust to occlusion. There are many occlusion cases in multi-object tracking datasets [28,11]. Even state-of-the-art detectors cannot detect objects under severe occlusions. Therefore, in many cases, an object will reappear after being occluded for a few frames. The key to getting a high IDF1 score is to preserve the identities of these reappeared objects, which is also a main challenge in multi-object tracking. Using appearance features is an effective way to preserve the identities of the occluded objects. However, we find in our experiments that the appearance feature distances of the same object increases linearly with the increase of the number of interval frames and the distance becomes even larger in the case of occlusion, as is shown in Figure 2. The feature distance becomes very large after 20 frames. Thus, it is difficult to retrieve the

lost object because it needs a high matching threshold, which may lead to other wrong matchings.

Another reason for the IDF1 performance gain of our method is that marginal probability is more robust to occlusion. We show some detailed visualization results of how our approach deals with occlusion and retrieve the lost object. As is shown in the second line of Figure 3, our approach can keep the identity unchanged in the case of severe occlusion. We adopt Kalman Filter to filter some impossible matchings (distance is infinity). The distance matrix in Figure 3 is the input of the Hungarian Algorithm in the first matching. The distance of the occluded person is 0.753 in the distance-based method and the distance is 0.725 in the probability-based method. From the distance distribution figure, we can see that the optimal matching threshold is 0.6 for the distanced-based method and 0.8 for the probability-based method. Only the probability-based method can preserve the identity of the occluded person as 0.725 is smaller than 0.8. Because we consider all possible matchings to compute the marginal probability, the probability of many impossible pairs can be very low (*e.g.* 0) and the distance is very large (*e.g.* 1) after using 1 to minus the probability. After adding the motion distance, the final distance is compressed between 0.8 to 1.2 and thus we can set a relatively high matching threshold (*e.g.* 0.8) and successfully retrieve the lost object with a large distance (*e.g.* 0.7). Also, by comparing 0.725 to 0.753, we can see that the marginal probability can make the distance of correct-matched occluded pairs lower.

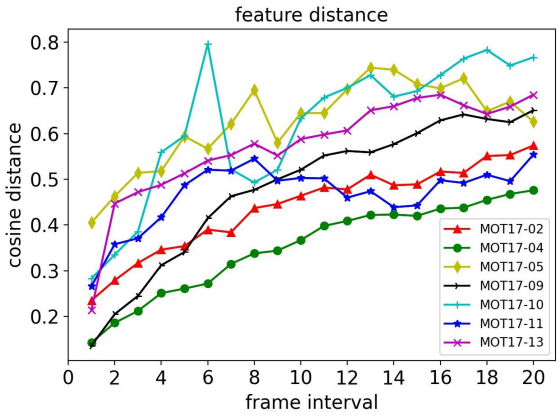

Fig. 2. Visualization of cosine distances of the appearance features at different frame intervals. The appearance features are extracted by the BoT Re-ID model. We show the results of all the sequences from the MOT17 training set.

4.4 Ablation studies

In this section, we compare different methods to compute marginal probability and evaluate different components of our matching strategy. We also evaluate the time-consuming of our method. We adopt Scaled-YOLOv4 [43] as the detector, BoT [26]

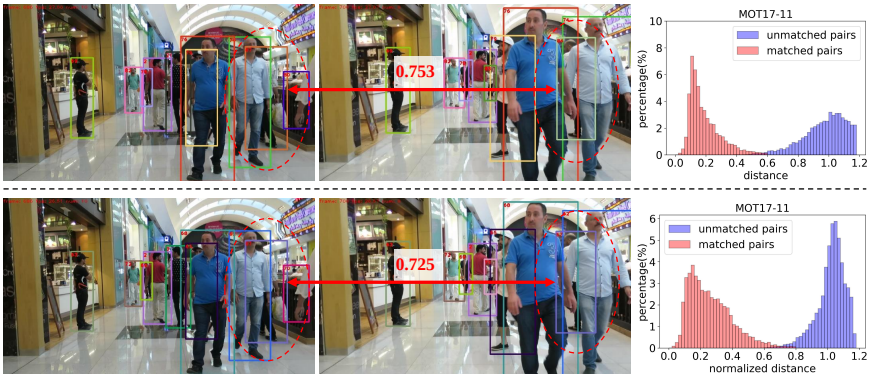

Fig. 3. Visualization of how our approach preserves object identity in the case of occlusion. The first line is the distanced-based method and the second line is the probability-base method. We show tracking visualizations, distance values, and distance distribution for both methods. The tracking results are from frame 686 and frame 700 of the MOT17-11 sequence. The occluded person is highlighted by the red dotted ellipse. The distance value is computed by the track and detection of the occluded person in frame 700. The occluded person is not detected from frame 687 to frame 699 and it is set as a lost track.

as the Re-ID model and evaluate on MOT17 training sets. We utilize powerful deep learning models and domain-different training data and evaluation data to make our setting as close to real-world applications as possible, which can better reflect the real performance of our method.

Probability computation. We compare different methods to compute the marginal probability, including softmax, bi-directional softmax, and our marginal inference method. The softmax method is to compute softmax probability between each track and all the detections. Bi-directional softmax is to compute probability between each track and all the detections along with each detection and all the tracks. We average the two probabilities to get the final probability.

The results are shown in Table 2. We can see that the bi-directional softmax method has the highest MOTA and lowest ID switches while our method has the highest IDF1 score. The softmax-based methods only consider one-to-n matchings and lack global consideration. Our method can approximate all possible matchings and thus has global consideration. Our method has sightly more ID switches than bi-softmax because sometimes the ID will switch 2 times and then turn to be correct in some cases of severe occlusion. In such cases, the IDF1 score is still high and we argue that IDF1 score is more important.

Matching strategy. We evaluate the effect of different components in the matching strategy, including appearance features, sparse probability, Kalman Filter and IoU. As is shown in Table 3, appearance and motion cues are complementary. IoU matching often works when appearance features are unreliable (*e.g.* occlusion). Finally, the marginal probability matching further increase the IDF1 score by about 1 point.

Method	MOTA↑	IDF1↑	IDs↓
Distance	71.5	74.8	499
Softmax	71.6	73.7	477
Bi-softmax	71.6	74.7	**405**
Marginal (Ours)	**71.6**	**75.7**	449

Table 2. Comparison of different methods to compute probability. "Bi-softmax" is short for bi-directional softmax. "Marginal" is short for marginal inference.

A	K	IoU	P	MOTA↑	IDF1↑	IDs↓
✓				68.8	71.9	792
✓	✓			70.1	73.6	777
✓	✓	✓		71.5	74.8	499
✓	✓	✓	✓	**71.6**	**75.7**	**449**

Table 3. Ablation study of different components in the matching strategy. "A" is short for appearance features, "P" is short for probability, "K" is short for Kalman Filter.

Time-consuming. We compute the time-consuming of the linking step using videos with different density (average number of pedestrians per frame). We compare the distance-based matching method to the probability-based matching method. We choose videos with different density from the MOT17 training set. As is shown in Figure 4, it takes only a fraction of time (less than 10 ms) to compute the marginal probability.

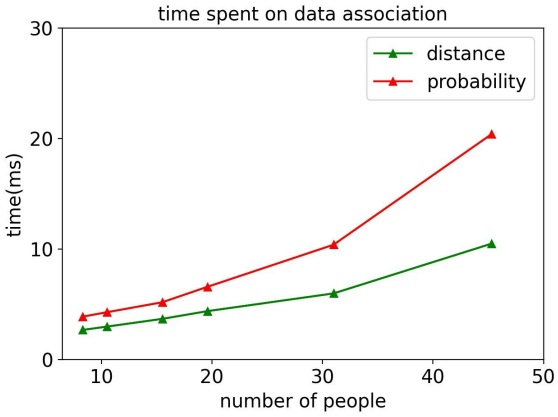

Fig. 4. Visualization of the time-consuming of data association. We evaluate two different methods on the training set of MOT17.

Public Detection								
Mode	Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
Off	MHT_DAM [18]	50.7	47.2	491	869	22875	252889	2314
Off	jCC [17]	51.2	54.5	493	872	25937	247822	1802
Off	FWT [15]	51.3	47.6	505	830	24101	247921	2648
Off	eHAF [39]	51.8	54.7	551	893	33212	236772	1834
Off	TT [55]	54.9	63.1	575	897	20236	233295	1088
Off	MPNTrack [8]	58.8	61.7	**679**	**788**	17413	213594	**1185**
Off	Lif_T [16]	**60.5**	**65.6**	637	791	**14966**	**206619**	1189
On	MOTDT [9]	50.9	52.7	413	841	24069	250768	2474
On	FAMNet [10]	52.0	48.7	450	787	14138	253616	3072
On	DeepMOT [50]	53.7	53.8	458	861	11731	247447	1947
On	Tracktor++v2 [2]	56.3	55.1	498	831	**8866**	235449	1987
On	CenterTrack [60]	61.5	59.6	621	752	14076	200672	2583
On	MTracker (Ours)	**62.1**	**65.0**	**657**	**730**	24052	**188264**	**1768**
Private Detection								
Mode	Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
On	TubeTK [30]	63.0	58.6	735	468	27060	177483	4137
On	CTracker [32]	66.6	57.4	759	570	22284	160491	5529
On	CenterTrack [60]	67.8	64.7	816	579	**18489**	160332	**3039**
On	FairMOT [58]	73.7	72.3	1017	408	27507	117477	3303
On	PermaTrackPr [41]	73.8	68.9	1032	405	28998	115104	3699
On	TransTrack [40]	75.2	63.5	1302	**240**	50157	86442	3603
On	CorrTracker [44]	76.5	73.6	1122	300	29808	99510	3369
On	MTracker (Ours)	**77.3**	**75.9**	**1314**	276	45030	**79716**	3255

Table 4. Comparison of the state-of-the-art methods on MOT17 test sets. We report results under both public detection and private detection protocols.

4.5 Benchmark evaluation

We compare our Marginal Inference Tracker (MTracker) with the state-of-the-art methods on the test sets of MOT17 and MOT20 under both public detection and private detection protocols. We list the results of both online methods and offline methods for completeness. We only compare directly to the online methods for fairness. For public detection results, we adopt the one-shot tracker FairMOT [58] to jointly perform detection and Re-ID and follow CenterTrack [60] to use public detections to filter the tracklets with a more strict IoU distance. For private detection results, we adopt a more powerful detector Scaled-Yolov4 [43] and Re-ID model BoT [26].

Table 4 and Table 5 show our results on the test sets of MOT17 and MOT20. For public detection results, MTracker achieves high IDF1 score and low ID switches and outperforms the state-of-the-art methods by a large margin. On MOT17 test sets, the IDF1 score of MTracker is 5.4 points higher than CenterTrack and the ID switches are reduced by 30%. On MOT20 test sets, the IDF1 score of MTracker is 12.3 points higher than Tractor++v2 [2] and the ID switches are reduced by 70%. The high IDF1 score and low ID switches indicate that our method has strong identity preservation ability, which

Public Detection								
Mode	Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
Off	IOU19 [6]*	35.8	25.7	126	389	24427	319696	15676
Off	V-IOU [7]*	46.7	46.0	288	306	33776	261964	2589
Off	MPNTrack [8]	**57.6**	**59.1**	**474**	**279**	**16953**	**201384**	**1210**
On	SORT20 [4]	42.7	45.1	208	326	27521	264694	4470
On	TAMA [53]*	47.6	48.7	342	**297**	38194	252934	2437
On	Tracktor++ [2]*	51.3	47.6	313	326	16263	253680	2584
On	Tracktor++v2 [2]	52.6	52.7	365	331	**6930**	236680	1648
On	MTracker (Ours)	**55.6**	**65.0**	**444**	388	12297	**216986**	**480**
Private Detection								
Mode	Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
On	MLT [54]	48.9	54.6	384	274	45660	216803	**2187**
On	FairMOT [58]	61.8	67.3	**855**	**94**	103440	**88901**	5243
On	TransTrack [40]	65.0	59.4	622	167	27197	150197	3608
On	CorrTracker [44]	65.2	69.1	-	-	79429	95855	5183
On	MTracker (Ours)	**66.3**	**67.7**	707	146	**41538**	130072	2715

Table 5. Comparison of the state-of-the-art methods on MOT20 test sets. We report results under both public detection and private detection protocols. The methods denoted by * are the ones reported on CVPR2019 Challenge in which the videos and ground-truth are almost the same as MOT20.

reveals the advantages of the marginal probability. For the private detection results, we use the same training data as FairMOT and substantially outperforms it on both MOTA and IDF1 score.

5 Conclusion

We present an efficient and robust data association method for multi-object tracking by marginal inference. The obtained marginal probability can be regarded as "normalized distance" and is significantly more stable than the distances based on Re-ID features. Our probability-based data association method has several advantages over the classic distance-based one. First, we can use a single threshold for all videos thanks to the stable probability distribution. Second, we empirically find that marginal probability is more robust to occlusion. Third, our approach is general and can be applied to the existing state-of-the-art trackers [58,48] easily. We hope our work can benefit real applications where data distribution always varies significantly.

Acknowledgement This work was in part supported by NSFC (No. 61733007 and No. 61876212) and MSRA Collaborative Research Fund.

References

1. Bae, S.H., Yoon, K.J.: Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1218–1225 (2014) 3, 4
2. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV. pp. 941–951 (2019) 8, 13, 14
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008) 7
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP. pp. 3464–3468. IEEE (2016) 2, 3, 4, 14
5. Birkhoff, G.: Tres observaciones sobre el algebra lineal. Univ. Nac. Tucuman, Ser. A **5**, 147–154 (1946) 6
6. Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2017) 3, 14
7. Bochinski, E., Senst, T., Sikora, T.: Extending iou based multi-object tracking by visual information. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2018) 14
8. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6247–6257 (2020) 13, 14
9. Chen, L., Ai, H., Zhuang, Z., Shang, C.: Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2018) 3, 4, 13
10. Chu, P., Ling, H.: Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In: ICCV. pp. 6172–6181 (2019) 13
11. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv:2003.09003[cs] (Mar 2020), http://arxiv.org/abs/1906.04567, arXiv: 2003.09003 7, 9
12. Fang, Y., Yang, S., Wang, S., Ge, Y., Shan, Y., Wang, X.: Unleashing vanilla vision transformer with masked image modeling for object detection. arXiv preprint arXiv:2204.02964 (2022) 3
13. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2009) 7
14. He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: Fastreid: a pytorch toolbox for real-world person re-identification. arXiv preprint arXiv:2006.02631 (2020) 8
15. Henschel, R., Leal-Taixé, L., Cremers, D., Rosenhahn, B.: Fusion of head and full-body detectors for multi-object tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1428–1437 (2018) 13
16. Hornakova, A., Henschel, R., Rosenhahn, B., Swoboda, P.: Lifted disjoint paths with application in multiple object tracking. In: International Conference on Machine Learning. pp. 4364–4375. PMLR (2020) 13
17. Keuper, M., Tang, S., Andres, B., Brox, T., Schiele, B.: Motion segmentation & multiple object tracking by correlation co-clustering. IEEE transactions on pattern analysis and machine intelligence **42**(1), 140–153 (2018) 13
18. Kim, C., Li, F., Ciptadi, A., Rehg, J.M.: Multiple hypothesis tracking revisited. In: Proceedings of the IEEE international conference on computer vision. pp. 4696–4704 (2015) 13

19. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Transactions on information theory **47**(2), 498–519 (2001) 5
20. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955) 1, 4, 6, 7
21. Lacoste-Julien, S., Jaggi, M.: On the global linear convergence of frank-wolfe optimization variants. arXiv preprint arXiv:1511.05932 (2015) 6
22. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017) 3
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 8
24. Lin, W., Liu, H., Liu, S., Li, Y., Qi, G.J., Qian, R., Wang, T., Sebe, N., Xu, N., Xiong, H., et al.: Human in events: A large-scale benchmark for human-centric video analysis in complex events. arXiv preprint arXiv:2005.04490 (2020) 8
25. Lu, Z., Rathod, V., Votel, R., Huang, J.: Retinatrack: Online single stage joint detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14668–14678 (2020) 1, 3
26. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 8, 10, 13
27. Martins, A., Astudillo, R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In: ICML. pp. 1614–1623. PMLR (2016) 6
28. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 7, 9
29. Niculae, V., Martins, A., Blondel, M., Cardie, C.: Sparsemap: Differentiable sparse structured inference. In: ICML. pp. 3799–3808. PMLR (2018) 6
30. Pang, B., Li, Y., Zhang, Y., Li, M., Lu, C.: Tubetk: Adopting tubes to track multi-object in a one-step training model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6308–6318 (2020) 13
31. Pang, J., Qiu, L., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense instance similarity learning. arXiv preprint arXiv:2006.06664 (2020) 1, 3, 4
32. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. arXiv preprint arXiv:2007.14557 (2020) 13
33. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 3
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015) 3, 7
35. Rezatofighi, S.H., Milan, A., Zhang, Z., Shi, Q., Dick, A., Reid, I.: Joint probabilistic data association revisited. In: Proceedings of the IEEE international conference on computer vision. pp. 3047–3055 (2015) 4
36. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV. pp. 17–35. Springer (2016) 7, 8
37. Shan, C., Wei, C., Deng, B., Huang, J., Hua, X.S., Cheng, X., Liang, K.: Fgagt: Flow-guided adaptive graph tracking. arXiv preprint arXiv:2010.09015 (2020) 1, 3, 4
38. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018) 8
39. Sheng, H., Zhang, Y., Chen, J., Xiong, Z., Zhang, J.: Heterogeneous association graph fusion for target association in multiple object tracking. IEEE Transactions on Circuits and Systems for Video Technology **29**(11), 3269–3280 (2018) 13

40. Sun, P., Jiang, Y., Zhang, R., Xie, E., Cao, J., Hu, X., Kong, T., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple-object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020) 1, 13, 14
41. Tokmakov, P., Li, J., Burgard, W., Gaidon, A.: Learning to track with object permanence. arXiv preprint arXiv:2103.14258 (2021) 13
42. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: Mots: Multi-object tracking and segmentation. In: CVPR. pp. 7942–7951 (2019) 3, 4
43. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Scaled-yolov4: Scaling cross stage partial network. arXiv preprint arXiv:2011.08036 (2020) 8, 9, 10, 13
44. Wang, Q., Zheng, Y., Pan, P., Xu, Y.: Multiple object tracking with correlation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3876–3886 (2021) 13, 14
45. Wang, Z., Zheng, L., Liu, Y., Wang, S.: Towards real-time multi-object tracking. arXiv preprint arXiv:1909.12605 (2019) 1, 3, 4
46. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 79–88 (2018) 8
47. Welch, G., Bishop, G., et al.: An introduction to the kalman filter (1995) 3
48. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017) 1, 2, 3, 4, 7, 8, 14
49. Xu, J., Cao, Y., Zhang, Z., Hu, H.: Spatial-temporal relation networks for multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3988–3998 (2019) 3, 4
50. Xu, Y., Osep, A., Ban, Y., Horaud, R., Leal-Taixé, L., Alameda-Pineda, X.: How to train your deep multi-object tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6787–6796 (2020) 13
51. Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., Lu, H.: Towards grand unification of object tracking. In: ECCV (2022) 3
52. Yang, F., Choi, W., Lin, Y.: Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2129–2137 (2016) 7
53. Yoon, Y.C., Kim, D.Y., Song, Y.m., Yoon, K., Jeon, M.: Online multiple pedestrians tracking using deep temporal appearance matching association. Information Sciences (2020) 14
54. Zhang, Y., Sheng, H., Wu, Y., Wang, S., Ke, W., Xiong, Z.: Multiplex labeling graph for near-online tracking in crowded scenes. IEEE Internet of Things Journal **7**(9), 7892–7902 (2020) 14
55. Zhang, Y., Sheng, H., Wu, Y., Wang, S., Lyu, W., Ke, W., Xiong, Z.: Long-term tracking with deep tracklet association. IEEE Transactions on Image Processing **29**, 6694–6706 (2020) 13
56. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. arXiv preprint arXiv:2110.06864 (2021) 3
57. Zhang, Y., Wang, C., Wang, X., Liu, W., Zeng, W.: Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 3
58. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021) 1, 3, 4, 7, 8, 9, 13, 14
59. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015) 8

60. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision. pp. 474–490. Springer (2020) 1, 3, 8, 13
61. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019) 3