

Learning a Proposal Classifier for Multiple Object Tracking

Peng Dai¹ Renliang Weng² Wongun Choi² Changshui Zhang¹ Zhangping He²
Wei Ding²

¹Tsinghua University, Beijng, China. ²Aibee Inc

¹{daip2020, zcs}@mail.tsinghua.edu.cn

²{rlweng, wgchoi, zphe, weiding}@aibee.com

Abstract

The recent trend in multiple object tracking (MOT) is heading towards leveraging deep learning to boost the tracking performance. However, it is not trivial to solve the data-association problem in an end-to-end fashion. In this paper, we propose a novel proposal-based learnable framework, which models MOT as a proposal generation, proposal scoring and trajectory inference paradigm on an affinity graph. This framework is similar to the two-stage object detector Faster RCNN, and can solve the MOT problem in a data-driven way. For proposal generation, we propose an iterative graph clustering method to reduce the computational cost while maintaining the quality of the generated proposals. For proposal scoring, we deploy a trainable graph-convolutional-network (GCN) to learn the structural patterns of the generated proposals and rank them according to the estimated quality scores. For trajectory inference, a simple deoverlapping strategy is adopted to generate tracking output while complying with the constraints that no detection can be assigned to more than one track. We experimentally demonstrate that the proposed method achieves a clear performance improvement in both MOTA and IDF1 with respect to previous state-of-the-art on two public benchmarks. Our code is available at https://github.com/daip13/LPC_MOT.git.

1. Introduction

Tracking multiple objects in videos is an important problem in many application domains. Particularly, estimating humans location and their motion is of great interest in surveillance, business analytics, robotics and autonomous driving. Accurate and automated perception of their whereabouts and interactions with others or environment can help identifying potential illegal activities, understanding customer interactions with retail spaces, planning the pathway of robots or autonomous vehicles.

The ultimate goal of multiple object tracking (MOT) is to estimate the trajectory of each individual person as one

complete trajectory over their whole presence in the scene without having any contamination by the others. Much research is done in this domain to design and implement robust and accurate MOT algorithms in the past [7, 28, 47]. However, the problem still remains unsolved as reported in the latest results in various public benchmarks [14, 16, 18, 37]. The key challenges in MOT are mostly due to occlusion and scene clutter, as in any computer vision problem. Consider the case when two people (yellow and purple boxes in Figure 1) are walking together in a spatial neighborhood. At one point, both people are visible to the camera and recent object detection algorithms like [33, 42, 43], can easily detect them. When the two people become aligned along the camera axis, however, one is fully occluded by another, and later both become visible when one passes the other. Since the visual appearance may have subtle difference between the two targets due to various reasons like illumination, shading, similar clothing, etc, estimating the trajectory accurately without contamination (often called as identity transfer) remains as the key challenge. In more crowded scenes, such occlusion can happen across multiple peoples which pose significant troubles to any MOT algorithm. Moreover, the MOT problem naturally has an exponentially large search space for the solution ¹ which prohibits us from using complicated mechanisms.

Traditional approaches focus on solving the problem by employing various heuristics, hand-defined mechanisms to handle occlusions [8, 28]. Multiple Hypotheses Tracking (MHT [28]) is one of the earliest successful algorithms for MOT. A key strategy in MHT to handle occlusions is to delay data-association decisions by keeping multiple hypotheses active until data-association ambiguities are resolved. Network flow-based methods [8, 9] have recently become a standard approach for MOT due to their computational efficiency and optimality. In this framework, the data-association problem is modeled as a graph, where each

¹The tracking-by-detection approach, which is the de-facto framework in MOT domain, needs to solve the data-association problem given detections at each timestamp. The size of hypothesis space is exponential to the number of detections [28].

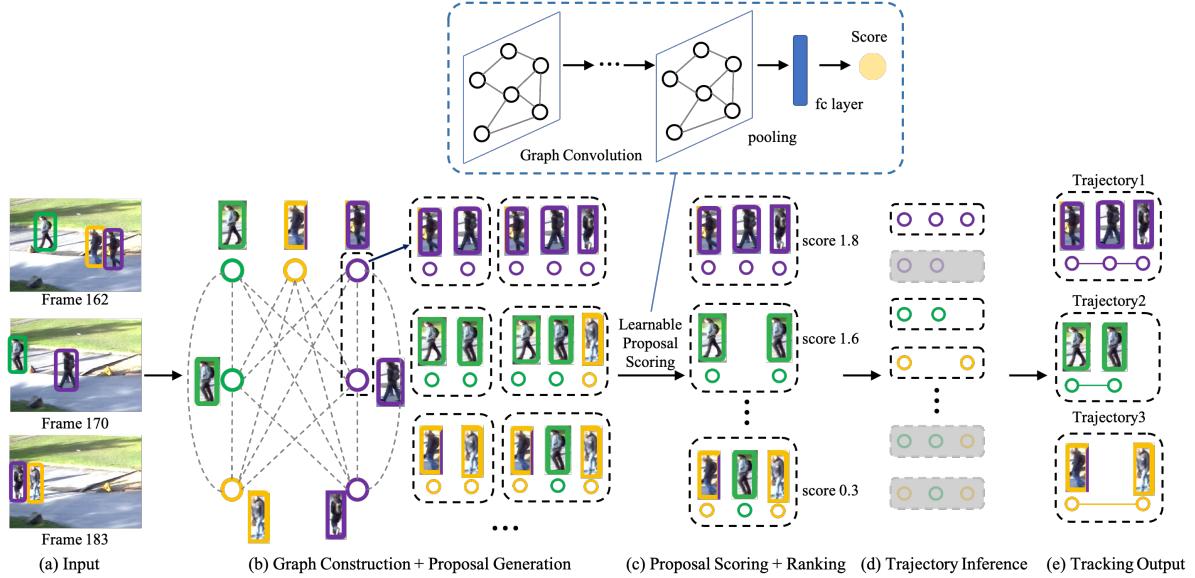


Figure 1. Overview of our framework. (a) Given a set of frames and detections as input. (b) A graph is constructed to model the data association problem. Nodes in the graph represent detections/tracklets and the edges indicate possible links among nodes. The nodes in different colors represent different objects. Similar to two-stage object detector faster RCNN, our method adopts a proposal-based framework. Multiple proposals (i.e., candidate object trajectories) are generated based on the affinity graph. (c) We evaluate the quality scores of the generated proposals with trainable GCN. (d) A simple de-overlapping strategy is adopted to do trajectory inference and (e) obtain the final tracking output.

node represents a detection and each edge indicates a possible link between nodes. Then, occlusions can be handled by connecting non-consecutive node pairs. Both MHT and network flow-based methods need to manually design appropriate gap-spanning affinity for different scenarios. However, it is infeasible to enumerate all possible challenging cases and to implement deterministic logic for each case.

In this paper, we propose a simple but surprisingly effective method to solve the MOT problem in a data-driven way. Inspired by the latest advancement in object detection [43] and face clustering [57], we propose to design the MOT algorithm using two key modules, 1) proposal generation and 2) proposal scoring with graph convolutional network (GCN) [30]. Given a set of short tracklets (locally grouped set of detections using simple mechanisms), our proposal generation module (see Figure 1(b)) generates a set of proposals that contains the complete set of tracklets for fully covering each individual person, yet may as well have multiple proposals with contaminated set of tracklets (i.e., multiple different people merged into a proposal). The next step is to identify which proposal is better than the others by using a trainable GCN and rank them using the learned ranking/scoring function (see Figure 1(c)). Finally, we adopt an inference algorithm to generate tracking output given the rank of each proposal (see Figure 1(d)), while complying with the typical tracking constraints like no detection assigned to more than one track.

The main contribution of the paper is in four folds: 1)

We propose a novel learnable framework which formulates MOT as a proposal generation, proposal scoring and trajectory inference pipeline. In this pipeline, we can utilize algorithms off the shelf for each module. 2) We propose an iterative graph clustering strategy for proposal generation. It can significantly reduce the computational cost while guaranteeing the quality of the generated proposals. 3) We employ a trainable GCN for proposal scoring. By directly optimizing the whole proposal score rather than the pairwise matching cost, GCN can incorporate higher-order information within the proposal to make more accurate predictions. 4) We show significantly improved state-of-the-art results of our method on two MOTChallenge benchmarks.

2. Related Work

Most state-of-the-art MOT works follow the tracking-by-detection paradigm which divides the MOT task into two sub-tasks: first, obtaining frame-by-frame object detections; second, linking the set of detections into trajectories. The first sub-task is usually addressed with object detectors [33, 42, 43, 56]. While the latter can be done on a frame-by-frame basis for online applications [22, 51, 53, 60, 61] or a batch basis for offline scenarios [3, 7, 38]. For video analysis tasks that can be done offline, batch methods are preferred since they can incorporate both past and future frames to perform more accurate association and are more robust to occlusions. A common approach to model

data-association in a batch manner is using a graph, where each node represents a detection and each edge indicates a possible link between nodes. Then, data-association can be converted to a graph partitioning task, i.e., finding the best set of active edges to predict partitions of the graph into trajectories. Specifically, batch methods differ in the specific optimization methods used, including network flow [41], generalized maximum multi clique [15], linear programming [24], maximum-weight independent set [8], conditional random field [55], k-shortest path [3], hyper-graph based optimization [50], etc. However, the authors in [4] showed that the significantly higher computational cost of these overcomplicated optimization methods does not translate to significantly higher accuracy.

As summarized in [12, 32], the research trend in MOT has been shifting from trying to find better optimization algorithms for the association problem to focusing on the use of deep learning in affinity computation. Most existing deep learning MOT methods focus on improving the affinity models, since deep neural networks are able to learn powerful visual and kinematic features for distinguishing the tracked objects from the background and other similar objects. Leal-Taixé et al. [31] adopted a Siamese convolutional neural network (CNN) to learn appearance features from both RGB images and optical flow maps. Amir et al. [46] employed long short-term memory (LSTM) to encode long-term dependencies in the sequence of observations. Zhu et al. [61] proposed dual matching attention networks with both spatial and temporal attention mechanisms to improve tracking performance especially in terms of identity-preserving metrics. Xu et al. [53] applied spatial-temporal relation networks to combine various cues such as appearance, location, and topology. Recently, the authors in [4, 45] confirmed the importance of learned re-identification (ReID) features for MOT. All aforementioned methods learn the pair-wise affinities independently from the association process, thus a classical optimization solver is still needed to obtain the final trajectories.

Recently, some works [7, 11, 47, 54] incorporate the optimization solvers into learning. Chu et al. [11] proposed an end-to-end model, named FAMNet, to refine feature representation, affinity model and multi-dimensional assignment in a single deep network. Xu et al. [54] presented a differentiable Deep Hungarian Net (DHN) to approximate the Hungarian matching algorithm and provide a soft approximation of the optimal prediction-to-ground-truth assignment. Schulter et al. [47] designed a bi-level optimization framework which frames the optimization of a smoothed network flow problem as a differentiable function of the pairwise association costs. Brasó et al. [7] modeled the non-learnable data-association problem as a differentiable edge classification task. In this framework, an undirected graph is adopted to model the data-association problem. Then, feature learn-

ing is performed in the graph domain with a message passing network. Next, an edge classifier is learned to classify edges in the graph into active and non-active. Finally, the tracking output is efficiently obtained via grouping connected components in the graph. However, this pipeline does not generally guarantee the *flow conservation constraints* [1]. The final tracking performance might be sensitive to the percentage of flow conservation constraints that are satisfied.

Similar to [7], our method also models the data-association problem with an undirected graph. However, our approach follows a novel proposal-based learnable MOT framework, which is similar to the two-stage object detector Faster RCNN [43], i.e. proposal generation, proposal scoring and proposal pruning.

3. Method

Given a batch of video frames and corresponding detections $\mathcal{D} = \{d_1, \dots, d_k\}$, where k is the total number of detections for all frames. Each detection is represented by $d_i = (o_i, p_i, t_i)$, where o_i denotes the raw pixels of the bounding box, p_i contains its 2D image coordinates and t_i indicates its timestamp. A trajectory is defined as a set of time-ordered detections $\mathcal{T}_i = \{d_{i_1}, \dots, d_{i_{n_i}}\}$, where n_i is the number of detections that form trajectory i . The goal of MOT is to assign a track ID to each detection, and form a set of m trajectories $\mathcal{T}_* = \{\mathcal{T}_1, \dots, \mathcal{T}_m\}$ that best maintains the objects' identities.

3.1. Framework Overview

As shown in Figure 1, our framework consists of four main stages.

Data Pre-Processing. To reduce the ambiguity and computational complexity in proposal generation, a set of tracklets $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_n\}$ is generated by linking detections \mathcal{D} in consecutive frames. And these tracklets \mathcal{T} are utilized as basic units in downstream modules.

Proposal Generation. As shown in Figure 1(b), we adopt a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} := \{v_1, \dots, v_n\}$, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, to represent the structured tracking data \mathcal{T} . A proposal $\mathcal{P}_i = \{v_i\}$ is a subset of the graph \mathcal{G} . The objective of proposal generation is to obtain an over-complete set of proposals which contain at least one perfect proposal for each target. However, it is computationally prohibitive to explore all perfect proposals $\{\hat{\mathcal{P}}_i\}_{i=1}^m$ from the affinity graph \mathcal{G} . Inspired by [57], we propose an iterative graph clustering strategy in this paper. By simulating the bottom-up clustering process, it can provide a good trade-off between proposal quality and the computational cost.

Proposal Scoring. With the over-complete set of proposals $\mathcal{P} = \{\mathcal{P}_i\}$, we need to calculate their quality scores and rank them, in order to select the subset of proposals that best represent real tracks. Ideally, the quality score can be

defined as a combination of precision and recall rates.

$$score(\mathcal{P}_i) = rec(\mathcal{P}_i) + w \cdot prec(\mathcal{P}_i) \quad (1)$$

$$rec(\mathcal{P}_i) = \frac{|\mathcal{P}_i \cap \hat{\mathcal{P}}_i|}{|\hat{\mathcal{P}}_i|} \quad (2)$$

$$prec(\mathcal{P}_i) = \begin{cases} 1, & \text{if } n(\mathcal{P}_i) = 1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where w is a weighting parameter controlling the contribution of precision score, $\hat{\mathcal{P}}_i$ is the ground-truth set of all detections with label $major(\mathcal{P}_i)$, and $major(\mathcal{P}_i)$ is the majority label of the proposal \mathcal{P}_i , $|\cdot|$ measures the number of detections, $n(\mathcal{P}_i)$ represents the number of labels included in proposal \mathcal{P}_i . Intuitively, $prec$ measures the purity, and rec reflects how close \mathcal{P}_i is to the matched ground-truth $\hat{\mathcal{P}}_i$. Inspired by [57], we adopt a GCN based network to learn to estimate the proposal score given the above definition. The precision of a proposal can be learned with a binary-cross-entropy loss through training procedure. However, it is much harder for a GCN to learn the recall of a proposal without exploring the entire graph structure including the vertices that are very far from a given proposal. We find that the normalized track length ($|\mathcal{P}_i|/C$, where C is a constant for normalization) is positively correlated with the recall of a proposal when precision is high. Thus, we approximate the recall rate of a proposal with the normalized track length and let the network to focus on accurately learning the precision of a proposal.

Trajectory Inference: Similar to the Non-Maximum Suppression in object detection, a trajectory inference strategy is needed to generate the final tracking output \mathcal{T}_* with the ranked proposals. This step is to comply with the tracking constraints like no tracklet assigned to more than one track. To reduce the computational cost, we adopt a simple de-overlapping algorithm with a complexity of $O(n)$.

3.2. Data Pre-processing

A tracklet is widely used as an intermediate input in many previous works [13, 58]. In our framework, we also use tracklets $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_n\}$ as basic units for graph construction, where n is the number of tracklets and is far less than detections k . Hence, it can significantly reduce overall computation. First, the ReID features a_i for each detection d_i is extracted with a CNN. Then, the overall affinity of two detections or detection-to-tracklet is computed by accumulating three elementary affinities based on their appearance, timestamps and positions. Finally, low-level tracklets are generated by linking detections based on their affinities with Hungarian algorithm [39]. It is worth noting that the purity of the generated tracklets is crucial, because the downstream modules use them as basic units and there is no strategy to recover from impure tracklets. Similarly

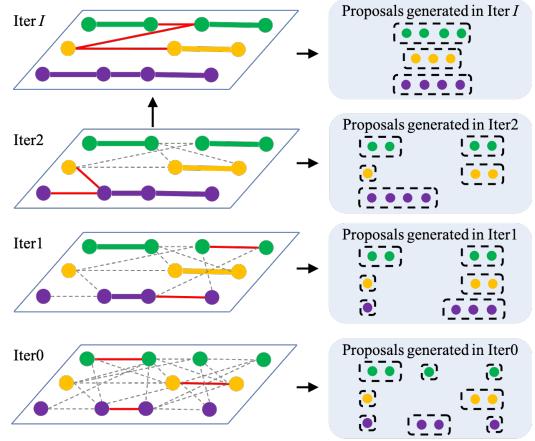


Figure 2. Visualization of the iterative proposal generation. In each iteration, only a small part of edges (red solid line) that meet the gating thresholds can be active. Each cluster generated in iteration i will be grouped as a vertex in iteration $i + 1$. To keep the purity of the clusters, strict gating thresholds are set in the first few iterations. As iterations increase, these thresholds will be gradually relaxed to grow proposals.

to [23], we use a dual-threshold strategy in which a higher threshold θ_1 is used to accept only associations with high affinities, and a lower threshold θ_2 is to avoid associations that have rivals with comparable affinities.

3.3. Iterative Proposal Generation

We propose an iterative clustering strategy to grow the proposals gradually, as shown in Figure 2. It mainly consists of two modules.

Affinity Graph Construction. At each iteration i , we build an affinity graph \mathcal{G} to model the similarity between vertices $\mathcal{V} := \{v_1, \dots, v_n\}$. Let vertex $v_i = (\mathbf{a}_i, \mathbf{t}_i, \mathbf{p}_i)$, where \mathbf{a}_i be the averaged ReID feature of a proposal, $\mathbf{t}_i = [t_i^s, \dots, t_i^e]$ be the sorted timestamps of detections in the proposal, $\mathbf{p}_i = [p_i^s, \dots, p_i^e]$ be the corresponding 2D image coordinates. The affinity score of an edge (v_i, v_j) is defined as the average score based on temporal, spatial and appearance similarities.

$$a_{ij}(v_i, v_j) = \frac{1}{3}(s_{ij}^a(\mathbf{a}_i, \mathbf{a}_j) + s_{ij}^t(\mathbf{t}_i, \mathbf{t}_j) + s_{ij}^p(\mathbf{p}_i, \mathbf{p}_j)) \quad (4)$$

$$s_{ij}^a(\mathbf{a}_i, \mathbf{a}_j) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{|\mathbf{a}_i| \cdot |\mathbf{a}_j|} \quad (5)$$

$$s_{ij}^t(\mathbf{t}_i, \mathbf{t}_j) = \begin{cases} \exp(-\frac{g(\mathbf{t}_i, \mathbf{t}_j)}{\sigma_t}), & \text{if } g(\mathbf{t}_i, \mathbf{t}_j) > 0 \\ -inf, & \text{otherwise} \end{cases} \quad (6)$$

$$s_{ij}^p(\mathbf{p}_i, \mathbf{p}_j) = \exp(-\frac{f(\mathbf{p}_i, \mathbf{p}_j)}{\sigma_p}) \quad (7)$$

where $g(\cdot)$ measures the minimum time gap between two vertices and $g(\mathbf{t}_i, \mathbf{t}_j) = -1$ if vertex v_i has temporal overlap.

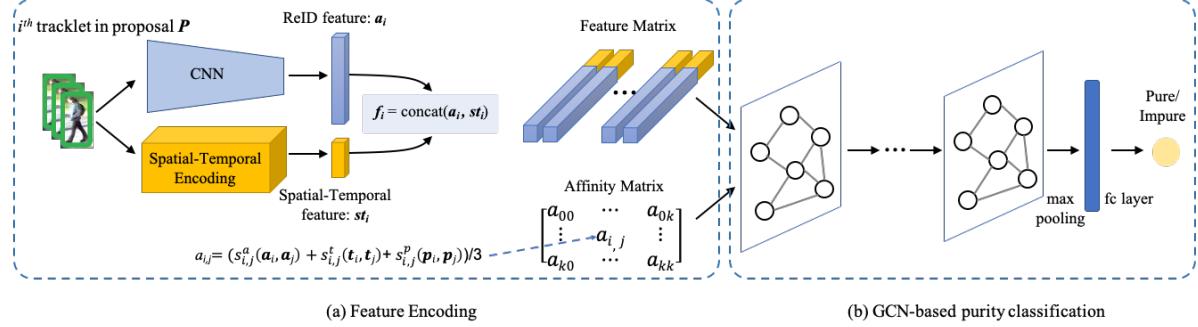


Figure 3. Visualization of (a) feature encoding and (b) GCN-based purity classification netowrk.

ping with vertex v_j , $f(\cdot)$ measures the Euclidean distance between the predicted box ² center of vertex v_i and the starting box center of vertex v_j , σ_t and σ_p are controlling parameters. To reduce the complexity of the graph, a simple gating strategy is adopted (see Appendix A.1 for details) and the maximum number of edges linked to one vertex is set to be less than K .

Cluster Proposals. The basic idea of proposal generation is to use connected components to find clusters. In order to keep the purity of the generated clusters high in the early iterations, we constrain the maximum size of each cluster to be below a threshold s_{max} . In this phase, the vertices of a target object may be over-fragmented into several clusters. The clusters generated in iteration i are used as the input vertices of the next iteration. And a new graph can be built on top of these clusters, thereby producing clusters of larger sizes. The final proposal set includes all the clusters in each iteration, thus providing an over-complete and diverse set of proposals $\mathcal{P} = \{\mathcal{P}_i\}$. The exact procedures are detailed in Algorithm 1 and 2 in Appendix A.2.

3.4. Purity Classification Network

In this subsection, we devise the purity classification network to estimate the precision scores $\{prec(\mathcal{P}_i)\}$ of the generated proposals \mathcal{P} . Specifically, given a proposal $\mathcal{P}_i = \{v_i\}_{i=1}^{N_i}$ with N_i vertices, the GCN takes the features associated with its vertices and sub-graph affinity matrix as input and predicts the probability of \mathcal{P}_i being pure. As shown in Figure 3, this module consists of the following two main parts.

Design of Feature Encoding. Both the appearance and the spatial-temporal features are crucial cues for MOT. For appearance features, a CNN is applied to extract feature embeddings a_i directly from RGB data of each detection d_i . Then, we obtain v_i 's corresponding appearance features \mathbf{a}_i by taking the average value of all detection appearance features. For spatial-temporal features, we seek to obtain a

²We apply a global constant velocity model to predict the 2D image coordinates of the bounding box.

representation that encodes, for each pair of temporal adjacent tracklets, their relative position, relative box size, as well as distance in time. For proposal $\mathcal{P}_i = \{v_i\}_{i=1}^{N_i}$, its vertices are sorted first in ascending order according to the start timestamp of each vertex. Then, for every pair of temporal adjacent tracklets v_i and v_{i+1} , the ending timestamp of v_i and the starting timestamp of v_{i+1} is denoted as t_{e_i} and $t_{s_{i+1}}$ respectively. And their bounding box coordinates in these timestamps are parameterized by top left corner image coordinates, width and height, i.e., (x_i, y_i, w_i, h_i) and $(x_{i+1}, y_{i+1}, w_{i+1}, h_{i+1})$. We compute the spatial-temporal feature \mathbf{st}_i for vertex v_i as:

$$\left(\frac{2(x_{i+1} - x_i)}{w_i + w_{i+1}}, \frac{2(y_{i+1} - y_i)}{h_i + h_{i+1}}, \log \frac{h_{i+1}}{h_i}, \log \frac{w_{i+1}}{w_i}, t_{s_{i+1}} - t_{e_i} \right) \quad (8)$$

if $i > 0$ else $\mathbf{st}_i = (1, 0, 0, 0, 0)$. With appearance feature \mathbf{a}_i and spatial-temporal feature \mathbf{st}_i at hand, we concatenate them to form the feature encoding $\mathbf{f}_i = concat(\mathbf{a}_i, \mathbf{st}_i)$ for each vertex v_i .

Design of GCN. As described above, we have obtained the features associated to vertices in \mathcal{P}_i (denoted as $\mathbb{F}_0(\mathcal{P}_i)$). As for the affinity matrix for \mathcal{P}_i (denoted as $\mathbb{A}(\mathcal{P}_i)$), a fully-connected graph is adopted, in which we compute the affinity between each pair of vertices, as shown in Figure 3 (a). The GCN network consists of L layers and the computation of each layer can be formulated as:

$$\mathbb{F}_{l+1}(\mathcal{P}_i) = \sigma(\mathbb{D}(\mathcal{P}_i)^{-1} \cdot (\mathbb{A}(\mathcal{P}_i) + \mathbb{I}) \cdot \mathbb{F}_l(\mathcal{P}_i) \cdot \mathbb{W}_l) \quad (9)$$

where $\mathbb{D}(\mathcal{P}_i) = \sum_j \mathbb{A}_{ij}(\mathcal{P}_i)$ is the diagonal degree matrix. $\mathbb{F}_l(\mathcal{P}_i)$ indicates the feature embeddings of the l -th layer, \mathbb{W}_l represents the transform matrix, and σ is a non-linear activation function (*ReLU* in our implementation). At the top-level feature embedding $\mathbb{F}_L(\mathcal{P}_i)$, a max pooling is applied over all vertices in \mathcal{P}_i to provide an overall summary. Finally, a fully-connected layer is employed to classify \mathcal{P}_i into a pure or impure proposal. As shown in Equation 9, for each GCN layer, it actually does three things: 1) computes the weighted average of the features of each vertex and its neighbors; 2) transforms the features with \mathbb{W}_l ; 3) feeds

the transformed features to a nonlinear activation function. Through this formulation, the purity network can learn the inner consistency of proposal \mathcal{P}_i .

3.5. Trajectory Inference

With the purity inference results, we can obtain the quality scores of all proposals with Equation 1. A simple de-overlapping algorithm is adopted to guarantee that each tracklet is assigned one unique track ID. First, we rank the proposals in descending order of the quality scores. Then, we sequentially assign track ID to vertices in the proposals from the ranked list, and modify each proposal by removing the vertices seen in preceding ones. The detailed algorithm is described in Algorithm 3 in Appendix A.2.

4. Experiments

In this section, we first present an ablation study to better understand the behavior of each module in our pipeline. Then, we compare our methods to published methods on the MOTChallenge benchmarks.

4.1. Experimental Setup

4.1.1 Datasets and metrics

All experiments are done on the multiple object tracking benchmark MOTChallenge, which consists of several challenging pedestrian tracking sequences with frequent occlusions and crowded scenes. We choose two separate tracking benchmarks, namely MOT17 [37] and MOT20 [16]. These two benchmarks consist of challenging video sequences with varying viewing angle, size, number of objects, camera motion, illumination and frame rate in unconstrained environments. To ensure a fair comparison with other methods, we use the public detections provided by MOTChallenge, and preprocess them by first running [4]. This strategy is widely used in published methods [7, 34].

For the performance evaluation, we use the widely accepted MOT metrics [5, 52, 44], including Multiple Object Tracking Accuracy (MOTA), ID F1 score (IDF1), Mostly Track targets (MT), Mostly Lost targets (ML), False Positives (FP), False Negatives (FN), ID switches (IDs), etc. Among these metrics, MOTA and IDF1 are the most important ones, as they quantify two of the main aspects of MOT, namely, object coverage and identity preservation.

4.1.2 Implementation details

ReID Model. For the CNN network used to extract ReID features, we employ a variant of ResNet50, named ResNet50-IBN [36], which replaces batch norm layer with instance-batch-norm (IBN) layer. After global average pooling layer, a batch norm layer and a classifier layer is added. We use triplet loss and ID loss to optimize the model

weights. For the ablation study, we use the ResNet50-IBN model trained on two publicly available datasets: ImageNet [17] and Market1501 [59]. While for the final benchmark evaluation, we add the training sequences in MOT17 [37] and MOT20 [16] to finetune the ResNet50-IBN model. Note that using training sequences in the benchmark to finetune ReID model for the test sequences is a common practice among MOT methods [21, 29, 49].

Parameter Setting. In affinity graph construction, the parameter σ_t and σ_p is empirically set to 40 and 100, respectively. In proposal generation, the maximum iteration number is set to $I=10$, the maximum neighbors for each node is set to $K=3$, the maximum cluster size is set to $s_{max}=2$, and the cluster threshold step is set to $\Delta=0.05$. In trajectory inference, the weighting parameter w is set to 1 and $C=200$.

GCN Training. We use a GCN with $L=4$ hidden layers in our experiments. The GCN model is trained end-to-end with Adam optimizer, where weight decay term is set to 10^{-4} , β_1 and β_2 is set to 0.9 and 0.999, respectively. The batch size is set to 2048. We train for 100 iterations in total with a learning rate 10^{-3} . For data augmentation, we randomly remove detections to simulate missed detections. For the ablation study, the leave-one-out cross-validation strategy is adopted to evaluate the GCN model.

Post Processing. We perform simple bilinear interpolation along missing frames to fill gaps in our trajectories.

4.2. Ablation Study

In this subsection, we aim to evaluate the performance of each module in our framework. We conduct all of our experiments with the training sequences of the MOT17 datasets.

4.2.1 Proposal Generation

To evaluate the performance of proposal generation, we choose the oracle purity network for proposal purity classification, i.e., determine whether the proposal \mathcal{P}_i is pure or not by comparing it with the ground-truth data. For baseline, we adopt the MHT algorithm [28] by removing the N -scan pruning step. To reduce the search space, a simple gating strategy is adopted which limits the maximum number of linkage for each vertex to be less than 20. The comparison results are summarized in Table 1. As expected, the time cost of our iterative proposal generation method is far less than that of the MHT-based method. Meanwhile, our method can achieve comparable MOTA and IDF1 scores. This demonstrates its ability to reduce the computational cost while guarantee the quality of the generated proposals.

Effect of Maximum Iteration Number. There are four parameters in proposal generation, namely I , K , s_{max} and Δ . Experimental results show that the tracking performance is insensitive to K , s_{max} and Δ . The detailed results are shown in Appendix B. Intuitively, increasing the maximum

Alg.	MOTA \uparrow	IDF1 \uparrow	MT \uparrow	ML \downarrow	FP \downarrow	FN \downarrow	IDs \downarrow	Hz \uparrow
Ours	64.8	73.3	631	384	4006	113769	749	21.6
MHT	64.7	73.6	632	389	3767	114495	608	2.4

Table 1. Performance comparison with different proposal generation algorithms.

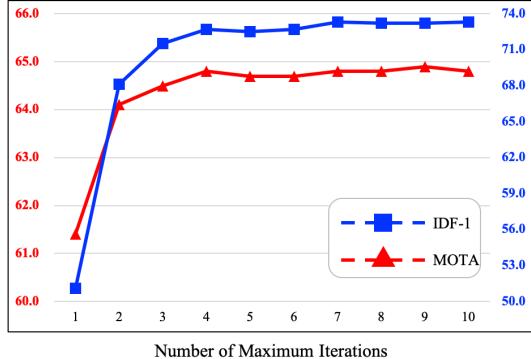


Figure 4. Influence of the iteration number I on proposal generation performance.

iteration number I allows to generate a larger number of proposals, and improves the possibility of the generated proposals to contain good tracklets under long-term occlusions. Hence, one would expect higher I values to yield better performance. We test this hypothesis in Figure 4 by doing proposal generation with increasing number of I , from 1 to 10. As expected, we see a clear upward tendency for both MOTA and IDF1 metrics. Moreover, it can be observed that the performance boost in both metrics mainly occurs when increasing I from 1 to 2, which demonstrates that most of the occlusions are short-term. We also observe that the upwards tendency for both MOTA and IDF1 metrics stagnates around seven iterations. There is a trade-off between performance and computational cost in choosing the proper number of iterations. Hence, we use $I = 10$ in our final configuration.

4.2.2 Purity Classification Network

Effects of the features. Our GCN-based purity classification network receives two main streams of features for each vertex: (i) appearance features from ReID model, and (ii) spatial-temporal features from Equation 8. We test their effectiveness by experimenting with combinations of the above two groups of features. Results are summarized in Table 2. It can be concluded that: (i) the appearance features seems to play a more important role in identity preservation, hence having higher IDF1 and MT measures, (ii) the spatial-temporal features can reduce the the number of FP and IDs, and (iii) combination of these two streams of features can improve the overall performance.

Feats.	MOTA \uparrow	IDF1 \uparrow	MT \uparrow	ML \downarrow	FP \downarrow	FN \downarrow	IDs \downarrow
Spat+Temp	63.6	69.4	622	381	5152	116653	819
App	64.0	70.3	634	373	6297	113865	1076
Spat+Temp+App	63.9	71.8	647	377	7176	113700	728

Table 2. Performance comparison for GCN-based purity classification network with different features.

Training Loss	MOTA \uparrow	IDF1 \uparrow	MT \uparrow	ML \downarrow	FP \downarrow	FN \downarrow	IDs \downarrow
BCELoss	63.9	71.8	647	377	7176	113700	728
MSELoss	63.8	71.2	646	378	7422	113878	765

Table 3. Performance comparison for GCN-based purity classification network with different loss functions.

Alg.	MOTA \uparrow	IDF1 \uparrow	MT \uparrow	ML \downarrow	FP \downarrow	FN \downarrow	IDs \downarrow
Oracle	64.8	73.3	631	384	4006	113769	749
GCN	63.9	71.8	647	377	7176	113700	728
TCN	63.8	70.6	628	379	6510	114666	901
ALSTM	63.5	69.5	634	380	6131	115756	1045
ALSTM-FCN	63.7	69.4	621	373	4897	116354	1087

Table 4. Performance comparison with different purity classification networks.

Effects of different loss functions. We perform an experiment to study the impact of different loss functions in model training. Table 3 lists the detailed quantitative comparison results by using binary-cross-entropy loss (BCELoss) and mean-squared-error loss (MSELoss), respectively. Using BCELoss shows a gain of 0.6 IDF1 measure and a small amount of decrease of IDs. Hence, we use BCELoss in our final configuration.

Effects of different networks. There are numerous previous works that use deep neural networks, such as Temporal Convolutional Network (TCN [2]), Attention Long-Short Term Memory (ALSTM [25]), ALSTM Fully Convolutional Network (ALSTM-FCN [25]) to conduct temporal reasoning on the sequence of observations. Table 4 presents the results by using these neural networks. It should be noticed that the oracle performance in Table 4 is obtained by using ground-truth data for purity classification. By comparing GCN with Oracle, we can see that GCN obtains better MT and ML measures, but worse MOTA and IDF1 measures than Oracle. The reason might be due to the false positives in GCN-based proposal purity classification, which would generate a few impure trajectories and hence reduce IDF1 measure. Moreover, the impure trajectories would cause quite a few FPs in the post processing (as shown in Table 4), hence reducing the MOTA measure. By comparing GCN with other neural networks, it is clear that GCN achieves better performance on most metrics, improving especially the IDF1 measure by 1.2 percentage. The performance gain is attributed to its capability of learning higher-

De-overlapping	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
Simple	63.9	71.8	647	377	7176	113700	728
Iterative Greedy	63.6	71.7	647	377	8628	113449	719

Table 5. Performance comparison with different de-overlapping strategies.

order information in a message-passing way to measure the purity of each proposal. It verifies that GCN is more suitable for solving the proposal classification problem.

4.2.3 Trajectory Inference

The iterative greedy strategy is a widely used technique in MOT, which can be an alternative choice of inference. Specifically, it iteratively performs the following steps: first, estimate the quality scores of all existing proposals; second, collect the proposal with highest quality score and assign unique track ID to the vertices within this proposal; third, modify the remaining proposals by removing the vertices seen in preceding ones. Hence, the computational complexity of the iterative greedy strategy is $O(N^2)$. Compared with the iterative greedy strategy, the simple de-overlapping algorithm only estimates the quality scores once. Therefore, it can reduce the computational complexity to $O(N)$. The comparison results are summarized in Table 5. It can be observed that the simple de-overlapping algorithm achieves slightly better performance in both MOTA and IDF1 metrics than the iterative greedy strategy. The reason might be due to that as the number of iteration increases, the number of nodes in each proposal decreases. Hence, the classification accuracy of the purity network might decrease.

4.3. Benchmark Evaluation

We report the quantitative results obtained by our method on MOT17 and MOT20 in Table 6 and Table 7 respectively, and compare it to methods that are officially published on the MOTChallenge benchmark. As shown in Table 6 and Table 7, our method obtains state-of-the-art results, improving especially the IDF1 measure by 1.2 percentage points on MOT17 and 3.4 percentage points on MOT20. It demonstrates that our method can achieve strong performance in identity preservation. We attribute this performance increase to our proposal-based learnable framework. First, our proposal generation module generates an over-complete set of proposals, which improves its anti-interference ability in challenging scenarios such as occlusions. Second, our GCN-based purity network directly optimizes the whole proposal score rather than the pairwise matching cost, which takes higher-order information into consideration to make globally informed predictions. We also provide more comparison results with other methods on MOT16 [37] benchmark in Appendix C.

Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓	Hz↑
Ours	59.0	66.8	29.9	33.9	23102	206948	1122	4.8
Lif.T [21]	60.5	65.6	27.0	33.6	14966	206619	1189	0.5
MPNTrack [7]	58.8	61.7	28.8	33.5	17413	213594	1185	6.5
JBNOT [20]	52.6	50.8	19.7	35.8	31572	232659	3050	5.4
eHAF [48]	51.8	54.7	23.4	37.9	33212	236772	1834	0.7
NOTA [10]	51.3	54.7	17.1	35.4	20148	252531	2285	-
FWT [19]	51.3	47.6	21.4	35.2	24101	247921	2648	0.2
jCC [27]	51.2	54.5	20.9	37.0	25937	247822	1802	1.8
GNNMatch [40]	57.3	56.3	24.2	33.4	14100	225042	1911	1.3
Tracktor [4]	56.3	55.1	21.1	35.3	8866	235449	1987	1.8
FAMNet [11]	52.0	48.7	19.1	33.4	14138	253616	3072	-

Table 6. Performance comparison with start-of-the art on MOT17 (top: offline methods; bottom: online methods).

Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓	Hz↑
Ours	56.3	62.5	34.1	25.2	11726	213056	1562	0.7
MPNTrack [7]	57.6	59.1	38.2	22.5	16953	201384	1210	6.5
GNNMatch [40]	54.5	49.0	32.8	25.5	9522	223611	2038	0.1
UnsupTrack [26]	53.6	50.6	30.3	25.0	6439	231298	2178	1.3
SORT20 [6]	42.7	45.1	16.7	26.2	27521	264694	4470	57.3

Table 7. Performance comparison with start-of-the art on MOT20.

Our method outperforms MPNTrack [7] only by a small margin in terms of the MOTA score. It should be noticed that MOTA measures the object coverage and overemphasizes detection over association [35]. We use the same set of detections and post-processing strategy (simple bilinear interpolation) as MPNTrack [7]. Then, achieving similar MOTA results is in line with expectations. IDF1 is preferred over MOTA for evaluation due to its focus on measuring association accuracy over detection accuracy. We also provide more qualitative results in Appendix D.

5. Conclusion

In this paper, we propose a novel proposal-based MOT learnable framework. For proposal generation, we propose an iterative graph clustering strategy which strikes a good trade-off between proposal quality and computational cost. For proposal scoring, a GCN-based purity network is deployed to capture higher-order information within each proposal, hence improving anti-interference ability in challenge scenarios such as occlusions. We experimentally demonstrate that our method achieves a clear performance improvement with respect to previous state-of-the-art. For future works, we plan to make our framework be trainable end-to-end especially for the task of proposal generation.

Acknowledgements. This research is funded by the National Key Research and Development Program of China (No. 2018AAA0100701)

References

[1] Ravindra K Ahyja, James B Orlin, and Thomas L Magnanti. *Network flows: theory, algorithms, and applications*. Prentice-Hall, 1993. 3

[2] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 7

[3] Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua. Multiple object tracking using k-shortest paths optimization. *PAMI*, 33(9):1806–1819, 2011. 2, 3

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *ICCV*, pages 941–951, 2019. 3, 6, 8

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468, 2016. 8

[7] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *CVPR*, pages 6247–6257, 2020. 1, 2, 3, 6, 8

[8] William Brendel, Mohamed Amer, and Sinisa Todorovic. Multiobject tracking as maximum weight independent set. In *CVPR*, pages 1273–1280, 2011. 1, 3

[9] Asad A Butt and Robert T Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *CVPR*, pages 1846–1853, 2013. 1

[10] Long Chen, Haizhou Ai, Rui Chen, and Zijie Zhuang. Aggregate tracklet appearance features for multi-object tracking. *IEEE Signal Processing Letters*, 26(11):1613–1617, 2019. 8

[11] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *ICCV*, pages 6172–6181, 2019. 3, 8

[12] Giuele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020. 3

[13] Peng Dai, Xue Wang, Weihang Zhang, and Junfeng Chen. Instance segmentation enabled hybrid data association and discriminative hashing for online multi-object tracking. *TMM*, 21(7):1709–1723, 2018. 4

[14] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, pages 436–454, 2020. 1

[15] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, pages 4091–4099, 2015. 3

[16] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1, 6

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[19] Roberto Henschel, Laura Leal-Taixé, Daniel Cremers, and Bodo Rosenhahn. Improvements to frank-wolfe optimization for multi-detector multi-object tracking. *arXiv preprint arXiv:1705.08314*, 2017. 8

[20] Roberto Henschel, Yunzhe Zou, and Bodo Rosenhahn. Multiple people tracking using body and joint detections. In *CVPRW*, pages 0–0, 2019. 8

[21] Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, and Paul Swoboda. Lifted disjoint paths with application in multiple object tracking. In *International Conference on Machine Learning*, pages 4364–4375, 2020. 6, 8

[22] Tao Hu, Lichao Huang, and Han Shen. Multi-object tracking via end-to-end tracklet searching and ranking. *arXiv preprint arXiv:2003.02795*, 2020. 2

[23] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, pages 788–801, 2008. 4

[24] Hao Jiang, Sidney Fels, and James J Little. A linear programming approach for multiple object tracking. In *CVPR*, pages 1–8, 2007. 3

[25] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. Multivariate lstm-fcns for time series classification. *Neural Networks*, 116:237–245, 2019. 7

[26] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. Simple unsupervised multi-object tracking. *arXiv preprint arXiv:2006.02609*, 2020. 8

[27] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *PAMI*, 42(1):140–153, 2018. 8

[28] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple hypothesis tracking revisited. In *ICCV*, pages 4696–4704, 2015. 1, 6

[29] Chanho Kim, Fuxin Li, and James M Rehg. Multi-object tracking with neural gating using bilinear lstm. In *ECCV*, pages 200–215, 2018. 6

[30] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[31] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *CVPRW*, pages 33–40, 2016. 3

[32] Laura Leal-Taixé, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, and Stefan Roth. Tracking the trackers: an analysis of the state of the art in multiple object tracking. *arXiv preprint arXiv:1704.02781*, 2017. 3

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 1, 2

[34] Qiankun Liu, Qi Chu, Bin Liu, and Nenghai Yu. Gsm: Graph similarity model for multi-object tracking. In *IJCAI*, pages 530–536, 2020. 6

[35] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, pages 1–31, 2020. 8

[36] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *TMM*, 22(10):2597–2609, 2019. 6

[37] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 6, 8

[38] Anton Milan, Konrad Schindler, and Stefan Roth. Multi-target tracking by discrete-continuous energy minimization. *PAMI*, 38(10):2054–2068, 2015. 2

[39] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957. 4

[40] Ioannis Papakis, Abhijit Sarkar, and Anuj Karpatne. Gcnmatch: Graph convolutional neural networks for multi-object tracking via sinkhorn normalization. *arXiv preprint arXiv:2010.00067*, 2020. 8

[41] Hamed Pirsiavash, Deva Ramanan, and Charles C Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, pages 1201–1208, 2011. 3

[42] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2, 3

[44] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016. 6

[45] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *CVPR*, pages 6036–6046, 2018. 3

[46] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *ICCV*, pages 300–311, 2017. 3

[47] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *CVPR*, pages 6951–6960, 2017. 1, 3

[48] Hao Sheng, Yang Zhang, Jiahui Chen, Zhang Xiong, and Jun Zhang. Heterogeneous association graph fusion for target association in multiple object tracking. *TCSVT*, 29(11):3269–3280, 2018. 8

[49] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, pages 3539–3548, 2017. 6

[50] Longyin Wen, Wenbo Li, Junjie Yan, Zhen Lei, Dong Yi, and Stan Z Li. Multiple target tracking based on undirected hierarchical relation hypergraph. In *CVPR*, pages 1282–1289, 2014. 3

[51] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649, 2017. 2

[52] Bo Wu and Ram Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *CVPR*, volume 1, pages 951–958, 2006. 6

[53] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *ICCV*, pages 3988–3998, 2019. 2, 3

[54] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. How to train your deep multi-object tracker. In *CVPR*, pages 6787–6796, 2020. 3

[55] Bo Yang and Ram Nevatia. An online learned crf model for multi-target tracking. In *CVPR*, pages 2034–2041, 2012. 3

[56] Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *CVPR*, pages 2129–2137, 2016. 2

[57] Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, and Dahua Lin. Learning to cluster faces on an affinity graph. In *CVPR*, pages 2298–2306, 2019. 2, 3, 4

[58] Haanju Yoo, Kikyung Kim, Moonsub Byeon, Younghan Jeon, and Jin Young Choi. Online scheme for multiple camera multiple target tracking based on multiple hypothesis tracking. *TCSVT*, 27(3):454–469, 2016. 4

[59] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 6

[60] Hui Zhou, Wanli Ouyang, Jian Cheng, Xiaogang Wang, and Hongsheng Li. Deep continuous conditional random fields with asymmetric inter-object constraints for online multi-object tracking. *TCSVT*, 29(4):1011–1022, 2018. 2

[61] Ji Zhu, Hua Yang, Nian Liu, Minyoung Kim, Wenjun Zhang, and Ming-Hsuan Yang. Online multi-object tracking with dual matching attention networks. In *ECCV*, pages 366–382, 2018. 2, 3