Spatial-Temporal Relation Networks for Multi-Object Tracking

Jiarui Xu[13]*, Yue Cao[23]*, Zheng Zhang[3], Han Hu[3]
[1]Hong Kong University of Science and Technology
[2]School of Software, Tsinghua University
[3]Microsoft Research Asia
jxuat@ust.hk, caoyue10@gmail.com, {zhez,hanhu}@microsoft.com

Abstract

Recent progress in multiple object tracking (MOT) has shown that a robust similarity score is key to the success of trackers. A good similarity score is expected to reflect multiple cues, e.g. appearance, location, and topology, over a long period of time. However, these cues are heterogeneous, making them hard to be combined in a unified network. As a result, existing methods usually encode them in separate networks or require a complex training approach. In this paper, we present a unified framework for similarity measurement between a tracklet and an object, which simultaneously encode various cues across time. We show a crucial principle to achieve this unified framework is the design of compatible feature representation for different cues and different sources (tracklet and object). A key technique behind this principle is a spatial-temporal relation module, which jointly models appearance and topology, and makes tracklet and object features compatible. The resulting method, named spatial-temporal relation networks (STRN), runs in a feed-forward way and can be trained in an end-to-end manner. The state-of-the-art accuracy was achieved on all of the MOT15∼17 benchmarks using public detection and online settings.

1. Introduction

Multiple object tracking (MOT) aims at locating objects and maintaining their identities across video frames. It has attracted a lot of attention because of its broad applications such as surveillance, sports game analysis, and autonomous driving. Most recent approaches follow the popular "tracking-by-detection" paradigm [12, 19, 27, 33, 35, 47, 58], where objects are firstly localized in each frame and then associated across frames. Such a decoupled pipeline reduces the overall complexity and shifts the major attention of MOT to a more unitary problem: object association.

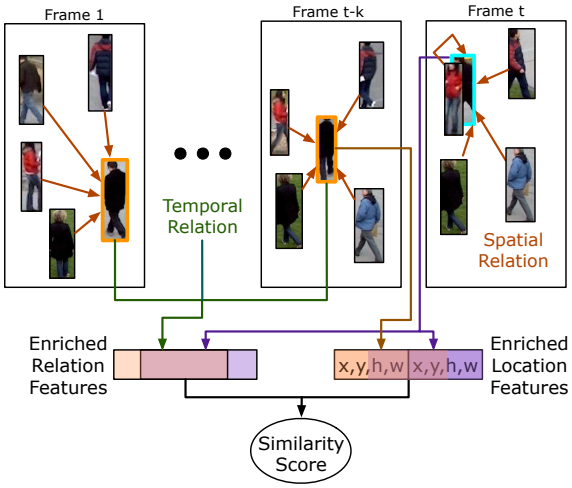

Figure 1. The proposed spatial-temporal relation networks (STRN) to compute similarity scores between tracklets and objects. The networks can combine various cues such as appearance, location, and topology, and aggregation information over time. The orange boxes and blue box indicate the same person in different frames.

This paradigm also benefits from the rapid progress in the field of object detection [15, 42, 60, 13] and has led several popular benchmarks for years, i.e. MOT15∼17 [28, 34].

In general, the performance of object association highly depends on a robust similarity score. The similarities in the most existing approaches are only based on the appearance features extracted from the cropped object patches [29]. The performance by such similarities is limited due to the following reasons: Firstly, the objects are often from the same category in tracking scenario, e.g. *person* in MOT15∼17 benchmark, with appearance hard to be distinguished. Secondly, objects across frames also suffer from frequent occlusions and pose variations, which further increases the difficulty in building a robust similarity score.

The pioneering works of exploring varying cues to build the similarity score have been proven to be effective[46, 12, 63, 58]. Convolutional neural networks have been well studied and employed to encode appearance cue [56, 63],

*This work is done when Jiarui Xu and Yue Cao are interns at Microsoft Research Asia.

and the hand-crafted location cues are integrated with appearance cue in recent works [46, 12, 63]. The topological structure [46] between bounding boxes is crucial for judging whether a pair of bounding boxes in different frames indicate the same object, especially for occlusion. As shown in Figure 1, the orange bounding boxes in frame 1 and frame $t-k$ and blue bounding box in frame t indicate the same person. Although the person in frame t is obscured by another person, and its appearance has a great difference compared with previous frames, the topological information keeps consistent and makes the obscured person identifiable. Besides, aggregation information across frames is also verified to be beneficial for measuring similarity [46, 26, 35].

However, because of the *heterogeneous* representation of different cues and resulting in the difficulties of dealing with all the cues into one unified framework, these works are usually based on cue-specific mechanisms [46, 26, 35, 26] and required sophisticated learning approaches [46]. For example, [46] uses an occupancy map to model topological information and [26] uses a specialized gating mechanism in RNN to aggregate information over time.

Our work is motivated by the success of relation networks in natural language problems [55] and vision problems [21, 57, 3, 48]. In the relation networks, each element aggregates features from other elements through a content-aware aggregation weight, which can be automatically learned according to the task goal without explicit supervision. Since there is not an excessive assumption about the data forms, the relation networks are widely used in modeling dependencies between distant, non-grid or differently distributed data, such as word-word relation [55], pixel-pixel relation [57] and object-object relation [21, 3, 48]. These data forms are hard to be modeled by regular convolution or sequential networks.

In this paper, we present a unified framework for similarity measurement by integrating multiple cues in an end-to-end manner through extending the object-object relation network [21] from the spatial domain to the spatial-temporal domain. With the extension of relation networks, we elegantly encode the appearance and topology cues for both objects and tracklets. The location-based cues are also made compatible by embedding the raw coordinates to a higher dimensional feature.

The whole module is illustrated in Figure 1. The approach is named spatial-temporal relation networks (STRN), which is fully feed-forward, can be trained in an end-to-end manner and achieves state-of-the-art performance over all online methods on MOT15$\sim$17 benchmarks.

2. Related Work

Tracking-by-Detection Paradigm Recent multiple object tracking (MOT) methods are mostly based on the tracking-by-detection paradigm, with the major focus on the object association problem. According to what kind of information is used to establishing the association between objects in different frames, the existing methods can be categorized into online methods [2, 12, 19, 27, 33, 35, 38, 47, 50, 58, 59, 9], and offline methods [7, 14, 36, 39, 40, 44, 51, 52, 53, 62]. The former methods are restricted to utilize past frames only in the association part, which is consistent with real-time applications. The latter methods can incorporate both past and future frames to perform more accurate association.

Our method also follows the tracking-by-detection paradigm and mainly focus on improving the measurement of object similarities. For better illustration and comparison with other methods, we only instantiate the online settings in this paper, but the proposed method is also applicable to both offline and online association.

Similarity Computation The major cues to compute similarities include appearance, location and topology.

The evolution of appearance feature extractor is from hand-craft [2, 38, 50, 59] to deep networks [63, 46, 12, 41, 26]. In this paper, we also utilize deep networks as our base appearance feature extractor. One crucial difference between the previous approaches lies in the way to build similarity from appearances. We utilize a hybrid of feature concatenation, cosine distance, location/motion priors to compute the final similarities.

The utilization of location/motion features is common as well. Most existing methods assume a prior motion model, such as slow velocity [5] and linear/non-linear motion model [63]. For example, the IoU trackers [5] rely on the assumption that objects in consecutive frames are expected to have high overlap, which is often not hold by fast moving objects. Other hard motion models also face the similar problem resulting in limited application. In this paper, instead of using hard location/motion priors, we integrate both unary location and motion information and learn the soft location/motion representation from data. Empirical studies on several benchmarks have proved the effectiveness of the learnable location representation.

The topological information is also crucial for measuring similarity [46]. However, leveraging such non-grid topology of multiple objects is challenging. Only a few works successfully encode the topological information, e.g. the occupancy grid in [46]. However, this occupancy grid only counts the distribution of objects, without differentiating individual objects. In this paper, we utilize relation networks to encode the topological information for making the individual object differentiable and identifiable.

Most existing methods utilize one or two cues for similarity computation, while only a few works trying to jointly learn all of them simultaneously [46]. Aggregating information across time [12, 46, 26, 63] is also rare. In addi-

tion, in order to learn the representations of different cues, these works usually adopt separate networks and sophisticated training strategy, e.g. a four-stage training in [46].

In this paper, we combine all of the mentioned cues across time for similarity measurement by using a unified framework, which is fully feed-forward and it can be trained in end-to-end. The key principle behind this unified framework is the compatible feature design between different cues (appearance, location and topology) and different sources (tracklet and object).

Relation Networks Recently, relation networks have been successfully applied in the fields of NLP, vision and physical system modeling [21, 55, 57, 3, 48], in order to capture long-term, non-grid and heterogeneous dependencies between elements.

Our approach is motivated by these works by extending the relation networks to multi-object tracking. In order to model the topological information of objects in the spatial domain and perform information aggregation over the temporal domain, we propose a spatial-temporal relation network. Although some recent works [12, 63] attempt to incorporate the attention mechanism into the multi-object tracking problem, they mainly aim at recovering salient foreground areas within a bounding box, thus alleviating the occlusion and ignoring the topology between objects.

3. Method

The goal of multi-object tracking (MOT) is to predict trajectories of multiple objects over time, denoted as $\mathbf{T} = \{\mathbf{T}_i\}_{i=1}^{N}$. The trajectory of the i^{th} object can be represented by a series of bounding boxes, denoted by $\mathbf{T}_i = \{\mathbf{b}_i^t\}_{t=1}^{T}, \mathbf{b}_i^t = [x_i^t, y_i^t, w_i^t, h_i^t]$. x_i^t and y_i^t denote the center location of the target i at frame t. w_i^t and h_i^t denote the width and height of the target object i, respectively.

Our method follows the online tracking-by-detection paradigm [58], which first detects multiple objects in each frame and then associates their identities across frames. The pipeline is illustrated in Figure 2. Given a new frame with the detected bounding boxes, the tracker computes similarity scores between the already obtained tracklets and the newly detected objects, resulting in a bipartite graph. Then the Hungarian algorithm [37] is adopted to get the optimal assignments. When running the assignment process frame-by-frame, object trajectories are yielded.

This paper is mainly devoted to building the robust similarity scores between tracklets extracted in previous frames and objects on the current frame, which proves crucial for multi-object tracking [29]. Denote the i^{th} tracklet before frame $t-1$ as $\mathbf{T}_i^{t-1} = \left\{\mathbf{b}_i^1, \mathbf{b}_i^2, ..., \mathbf{b}_i^{t-1}\right\}$ and the extracted objects at current frame t as $\mathbf{D}_t = \left\{\mathbf{b}_j^t\right\}_{j=1}^{N_t}$. Each pair $\left(\mathbf{T}_i^{t-1}, \mathbf{b}_j^t\right)$ is assigned a similarity score s_{ij}^t.

As mentioned before, the appearance, location, topology cues, and aggregating information over time are all useful in computing the similarity scores. In this paper, we present a novel method based on spatial-temporal relation networks to simultaneously encode all of the mentioned cues and perform reasoning across time. Figure 3 summarizes the entire process of similarity computation. Firstly, basic appearance features are extracted by a deep neural network, i.e. ResNet-50 in this paper, for both objects on current frame and objects on already obtained tracklets in previous frames, denoted as ϕ_i^t (object i on frame t). Then the appearance features of objects across space and time are reasoned through a spatial-temporal relation module (STRM), resulting in spatial strengthened representation and temporal strengthened representation, denoted as $\phi_{\text{S},i}^t$ and $\phi_{\text{ST},i}^t$, respectively. Through these two strengthened features, we further develop the two types of relationship features ϕ_R and ϕ_C by concatenating them together and calculating the cosine distance between them, respectively. Finally, we combine the relation features with the unary *location* feature ϕ_L and *motion* feature ϕ_M together as the representation of a tracklet-object pair $\left(\mathbf{T}_i^{t-k}, \mathbf{b}_j^t\right)$. Accordingly, the similarity is obtained by applying a two-layer network with a sigmoid function on the concatenation of the four features about a tracklet-object pair, ϕ_R, ϕ_C, ϕ_L and ϕ_M:

$$s_{ij}^t = \text{sigmoid}\left(W_{s2} \cdot \text{ReLU}(W_{s1} \cdot [\phi_R; \phi_C; \phi_L; \phi_M])\right), \quad (1)$$

A major difficulty of applying Eqn. (1) for the tracklet-object similarity computation lies in the incompatibility between different cues (appearance, location and topology) and between different sources (tracklet and object). We propose several techniques to tackle this difficulty: 1) the topology cue and appearance cues are jointly encoded through a spatial relation network; 2) the "cosine" distance Φ_C is introduced to facilitate the learning of compatible features; 3) a high dimensional *embedded* feature other than the raw coordinates is adopted for location/motion cues, for more compatible representation with other cues; 4) temporal reasoning is achieved by weighted averaging on features across frames, resulting in a tracklet feature compatible with that of a single object.

In the following subsections, we firstly introduce the spatial-temporal relation module (STRM), which corresponds to techniques 1) and 4) above. Then we present the

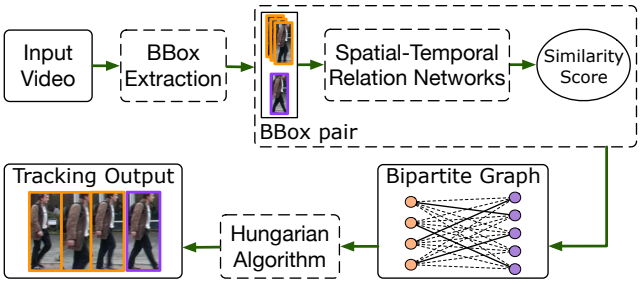

Figure 2. The online tracking-by-detection pipeline for multi-object tracking.

detailed design of the four features about a tracklet-object pair, ϕ_R, ϕ_C, ϕ_L and ϕ_M, following the "*increasing compatibility*" principle.

3.1. Spatial-Temporal Relation Module

We firstly review the basic *object relation module*, which is introduced in [21] to encode context information for object detection.

Object relation module (ORM) The basic object relation module [21] aims at strengthening an input appearance feature by aggregating information from other objects within a static image(a static image is a single frame in video). We denote object by $o_i = (\phi_i, \mathbf{b}_i)$, with ϕ_i the input appearance feature and $\mathbf{b}_i = (x_i, y_i, w_i, h_i)$ the object location. The object relation module computes a refined feature of object o_i by aggregating information from an object set $\mathcal{O} = \{o_j\}_{j=1}^N = \{(\phi_j, \mathbf{b}_j)\}_{j=1}^N$:

$$\phi_i' = \phi_i + \sum_j \omega_{ij} \cdot (W_V \cdot \phi_j), \tag{2}$$

where ω_{ij} is the attention weight contributed from object o_j to o_i; W_V is a transformation matrix of the input features.

Attention weight ω_{ij} is computed considering both the projected appearance similarity ω_{ij}^A and a geometric modulation term ω_{ij}^G as

$$\omega_{ij} = \frac{\omega_{ij}^G \cdot \exp\left(\omega_{ij}^A\right)}{\sum_{k=1}^N \omega_{ik}^G \cdot \exp\left(\omega_{ik}^A\right)}. \tag{3}$$

ω_{ij}^A is denoted as the scaled dot product of projected appearance features (W_Q, W_K are the projection matrices and d is the dimension of projected feature) [55], formulated as

$$\omega_{ij}^A = \frac{\langle W_Q \phi_i, W_K \phi_j \rangle}{\sqrt{d}}. \tag{4}$$

ω_{ij}^G is obtained by applying a small network on the relative location $\log\left(\frac{|x_i - x_j|}{w_j}, \frac{|y_i - y_j|}{h_j}, \frac{w_i}{w_j}, \frac{h_i}{h_j}\right)$. The original object relation module in [21] only performs reasoning within the spatial domain. In order to better leverage the advantage of object relation module in multi-object tracking, we extend this module to the temporal domain in this paper.

Extension to the spatial-temporal domain The object relation module can be extended to the spatial-temporal domain in a straight-forward way by enriching the object set $\mathcal{O}$ by all objects from previous frames. Such solution is obviously sub-optimal: firstly, the complexity is significantly increased due to more objects involved in reasoning; secondly, the spatial and temporal relations are tackled with no differentiation. In fact, spatial and temporal relations are generally expected to contribute differently to the encoding of cues. The spatial relation could draw on strengths in modeling *topology* between objects. The temporal relation is fit for aggregating information from multiple frames, which could potentially avoid the degradation problem caused by accidental low-quality bounding boxes.

Regarding the different effects of spatial and temporal relations, we present a separable spatial-temporal relation module, as illustrated in Figure 1. It firstly performs relation reasoning in the spatial domain on each frame. The spatial reasoning process strengthens input appearance features with automatically learned topology information. Then the strengthened features on multiple frames are aggregated through a temporal relation reasoning process.

The spatial relation reasoning process strictly follows equation 1 to encode topological clues, and the output characteristic of the process is expressed as p, whose encoded topological structure has been proved to be effective in the field of object detection.

The two types of relations follow different formulations. The spatial relation reasoning process strictly follows Eqn. (2) to encode the *topology* cue and the resulting output feature of this process is denoted as $\phi_{S,i}$, which has been proved to be effective in encoding the *topology* information to improve object detection [21]. Figure 5 illustrates the learnt spatial attention weights across frames. In general, the attention weights are stable on different frames, suggesting certainly captured topology representation. It should be noticed that the attention weight of an object itself is not necessarily higher than others, since W_Q and W_K in Eqn. (2) are different projections. This is also the case for geometric weights.

The temporal relation reasoning process is conducted right after spatial relation reasoning[1]. Instead of strengthening particular object features on each frame as in spatial relation modeling, we compute a representation of the whole tracklet by aggregating features from multiple frames. Due to the limiting of memory, the aggregation is only performed on latest τ_1 frames($\tau_1 = 10$ in default):

$$\phi_{ST,i}^t = \sum_{k=0}^{\tau_1 - 1} \omega_i^{t-k} \cdot \phi_{S,i}^{t-k}. \tag{5}$$

The attention weight is defined on the individual input feature as

$$\omega_i^t = \frac{\exp(\langle \mathbf{w}_T, \phi_{S,i}^t \rangle)}{\sum_k \exp(\langle \mathbf{w}_T, \phi_{S,i}^k \rangle)}. \tag{6}$$

where $\mathbf{w}_T$ is the weight vector of temporal relation. Eqn. (5) is essentially a weighted average of object features from recent frames. The learnt temporal attention weights is illustrates in Figure 4. The blurring, wrongly cropped or partly occluded detections are assigned with low attention weights, indicating feature qualities are automatically learnt, and the representation of a tracklet will be less affected by these low quality detections.

[1]Note the temporal relation reasoning is only performed for tracklets. The encoding of objects on current frame only includes spatial reasoning.

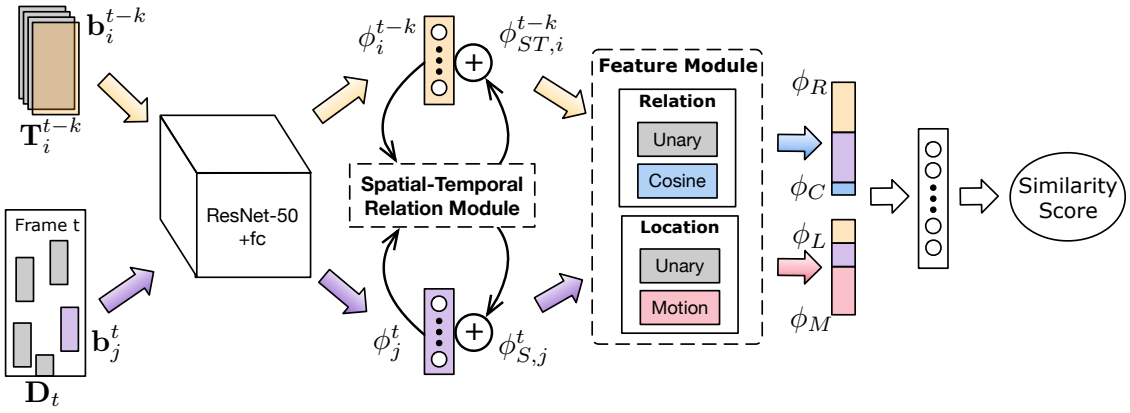

Figure 3. The architecture of Spatial-Temporal Relation Networks (STRN) to compute similarity scores between tracklets and objects.

Figure 4. Learnt temporal attention weights.

3.2. Tracklet-Object Pair Feature Design

As mentioned before, a key principle for the design of tracklet-object features (ϕ_R, ϕ_C, ϕ_L, ϕ_M) lies in making the representations of different cues (appearance, topology and location) and different sources (tracklet and object) compatible. While the previous subsection mainly discusses the joint encoding of the appearance and topology cues of an individual object or tracklet, this section presents the detailed feature design to represent the tracklet-object pair relationship, including the relation features ϕ_R, ϕ_C accounting for the similarity of appearance and topology cues, and the location/motion features ϕ_L, ϕ_M accounting for the location and motion relationship.

3.2.1 Relation Features

The spatial relation module couple the appearance cue and topology cue of an object. The temporal relation module aggregating information across frames.

Since an object corresponded tracklet may exceed the image boundary, or be lost tracked due to the imperfection of the system, the tracklet does not necessarily appear at last frame. We need to enlarge the candidate tracklets from the last frame to multiple frames. Because of the memory limiting, only recent τ_2 frames are involved(τ_2=10 in default).

We directly perform a linear transform on input relation features, which are regarded as the base feature type.

$$\phi_R = W_R \cdot \left[\phi^{t-k}_{ST,i}; \phi^{t}_{S,j}\right], 1 \le k \le \tau_2 \tag{7}$$

where W_R is a linear transform for feature fusion.

Directly using the concatenated relation features enables computing similarity of different modes. However, the freedom in representation is double-edged that it also increases the difficulty in learning compact individual features.

To address this issue, we propose to explicitly compute the cosine similarity between two relation features:

$$\phi_C = \cos\left(W_C \cdot \phi^{t-k}_{ST,i}, W_C \cdot \phi^{t}_{S,j}\right), 1 \le k \le \tau_2 \tag{8}$$

where W_C is a linear layer to project the original relation features into a low-dimensional representation, e.g. 128-d. Note we adopt shared W_C for both ϕ_{ST} and ϕ_S such that the two terms are "cosine" compatible, which performs better than using independent transformations for the two terms.

The cosine value is taken as an additional 1-d feature and fed to the following network for final similarity computation. The generation of hybrid relation features are summarized in Figure 6 (top).

In general, cosine value could take effect only in the scenarios where two input features are compatible in representation. At a first glance, it is not applicable to our "incompatible" features. Nevertheless, the features of tracklets and objects are actually compatible in some sensible way. The temporal relation in Eqn. (5) is basically a weighted average over features from multiple frames. There is no projection between the object feature and tracklet feature. Hence, they still locate at a close space and are suitable to be modeled by cosine value.

In the experiments, the hybrid representation of pair relation features achieves superior accuracy than the methods using each of the formulations alone.

3.2.2 Location Features

Location/motion feature is another widely used cues in building the similarity score. We take the location/motion

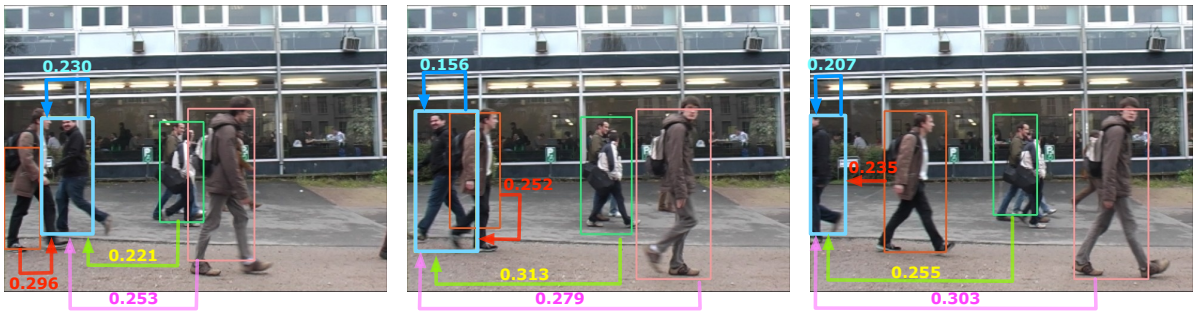

Figure 5. Learnt spatial attention weights across frames.

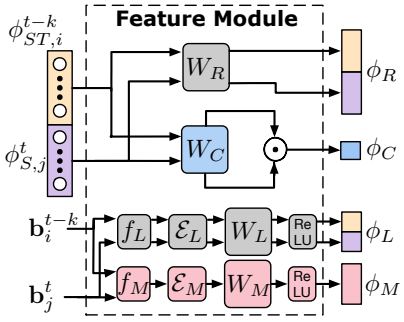

Figure 6. The design of tracklet-object feature representation. ϕ_R, ϕ_C, ϕ_L and ϕ_M indicate the relation feature, cosine similarity, location feature and motion feature, respectively. All the features will be concatenated to produce the final similarity scores through a two-layer network with a sigmoid function.

features from the last frame of a tracklet to represent the entire one, because the location/motion model in distant frames may drift a lot from the current frame.

The location features can be conveniently incorporated in our pipeline. The bare location features are firstly embedded and projected to the higher-dimensional space and then concatenated with the relation features to producing the final similarity score.

We embed and project of bare location features follow [55, 21] as

$$\phi_* = W_* \cdot \mathcal{E}_* \left(f_*(\mathbf{b}_i^{t-k}, \mathbf{b}_j^t) \right), \tag{9}$$

where $* \in \{L, M\}$ denotes the studied two types of location features, location and motion. The first one is the normalized absolute location of bounding box:

$$f'_L\left(\mathbf{b}_j^t\right) = \left(\frac{x_j^t}{I_w^t}, \frac{y_j^t}{I_h^t}, \frac{w_j^t}{I_w^t}, \frac{h_j^t}{I_h^t} \right), \tag{10}$$

where I_w^t and I_h^t are the width and height of frame t. f_L in Eqn 9 is defined as $f_L\left(\mathbf{b}_i^{t-k}, \mathbf{b}_j^t\right) = \left[f'_L(\mathbf{b}_i^{t-k}); f'_L(\mathbf{b}_j^t)\right]$.

The above location feature relates to the low-velocity assumption of objects, which has been proved work surprisingly well in the recent work [6]. Rather than using a hard constraint that the same objects on consecutive frames should have overlap, we incorporate the constraint softly into the feature representation, and the location patterns are learned from the data. The other location feature depict the motion information of an object in consecutive frames:

$$f_M\left(\mathbf{b}_i^{t-k}, \mathbf{b}_j^t\right) = \log\left(\frac{|x_i^{t-k} - x_j^t|}{kw_i^{t-k}}, \frac{|y_i^{t-k} - y_j^t|}{kh_i^{t-k}}, \frac{w_j^t}{kw_i^{t-k}}, \frac{h_j^t}{kh_i^{t-k}} \right). \tag{11}$$

This location (motion) feature relates to the constant velocity assumption of objects, which is proved as a effective information for a robust similarity score.

4. Experiments

4.1. Datasets and Evaluation Metrics

We utilize three MOT Benchmarks [28, 34] for evaluation. The benchmarks are challenging due to the large variety in frame rates, resolution, viewpoint, weather, camera motion and etc. These benchmarks are widely used in the field of multi-object tracking to evaluate different trackers.

2D MOT2015 consists of 11 training sequences and 11 testing sequences [28]. Following [46], we split the training sequences into two subset of 4 training and 6 validation sequences for ablation study.

MOT16 consists of 7 training sequences and 7 testing sequences. The scenes are mostly crowd pedestrians and are regarded as more challenging.

MOT17 use the same videos as the *MOT16* datasets but with better annotation and public detectors. All sequences are provided with three sets of detection results (DPM [15], Faster-RCNN [42] and SDP [60]) for more comprehensive comparison of different multi-object trackers.

For a fair comparison, we use the public detection result provided with datasets as the input of our approach.

Evaluation Metric We adopt the standard metrics of MOT Benchmarks [28, 34] for evaluation, including Multiple Object Tracking Accuracy (MOTA) [4], Multiple Object Tracking Precision (MOTP) [4], ID F1 Score (IDF1, the ratio of correctly identified detections over the average number of ground-truth and computed detections) [43], ID Precision (IDP, the fraction of detected identities correctly identified), [43], ID Recall (IDR, the fraction of ground truth identities correctly identified), [43], Mostly tracked targets (MT, the ratio of ground-truth trajectories covered by an output trajectory for at least 80% of ground truth

length), Mostly lost targets (ML, the ratio of ground-truth trajectories covered by an output trajectory for at most 20% of ground truth length), the number of False Positives (FP), the number of False Negatives (FN), the number of Identity Switch (IDS) [31], the number of Fragment Error (Frag). The latest Average Ranking (AR) on the MOT benchmark website is also reported, which is computed by taking the average of benchmark ranking of all metrics above.

4.2. Implementation Details

Network Architecture We use ResNet-50 [17] as our backbone network. We first train it on ImageNet Image Classification task [45] and then finetune the model on the MOT training datasets.

Given the bounding boxes of public detection, we crop and resize them to the resolution of 128×64. The cropped images are fed into the backbone network, producing a feature map with the resolution of 4×2. a new 256-d 1×1 convolution is applied on this feature map to reduce the channel dimension. A fully connected layer with dimension 1024 is applied right after the new 1×1 conv layer, which is used as the representing appearance feature ϕ_i (see Section 3.1).

In the spatial-temporal relation module, we mainly follow [21, 55] for the hyper-parameters of spatial relation reasoning. For temporal relation, the object features from the latest 9 frames are aggregated.

After the relation module, pairing relation features and location features are extracted. The linear layers W_R, W_c are of dimension 32 and 128, respectively. The function $\mathcal{E}_L$ embeds the 4-d bare location features to 64-d, followed by a linear layer W_L to project the feature to 16-d. All of the relation features and location features are concatenated, forming a 65-d feature and fed to a two-layer network with a sigmoid function.

Training During training, all detection bounding boxes in input frames are cropped and fed into the network. On average, each mini-batch contains 45 cropped images. A total of 437k, 425k and $1,275$k iterations are performed for 2DMOT2015, MOT16, MOT17 respectively. The learning rate is initialized as 10^{-3} and then decayed to 10^{-4} in the last $\frac{1}{3}$ training. Online hard example mining (OHEM) was adopted to address the heavy imbalance of positive/negative issue.

Inference In inference, the similarities between tracklets and objects on the current frame are computed according to Section 3.2. The association is then achieved by solving the bipartite graph as in Figure 2.

Following the common practice for online tracking approaches [58, 63, 12, 46], we consider the too short tracklets as false alarms. Specifically, for a sequence with the frame rate of F, we remove the short tracklets if it is matched less than $0.3F$ times in the past F frames after the initial match.

Feature	MOTA	MOTP	IDF	MT(%)	ML(%)	FP	FN	IDS
A_u	19.8	72.3	26.2	4.7	53.4	1,800	14,309	2,177
A_c	25.2	72.5	32.5	8.1	55.1	2,474	14,368	726
A	29.8	72.2	38.6	9.8	49.6	2,734	12,956	515
A+L_u	31.7	72.7	40.8	8.5	54.2	1,477	13,946	355
A+L_m	31.0	72.5	44.1	9.0	54.3	1,971	13,801	167
A+L	32.3	72.3	47.1	8.1	52.6	2,004	13,496	129

Table 1. Ablation study of various design of feature representation. A_u and L_u denote concatenation of individual relation and location features. A_c and L_m stand for cosine appearance feature and motion feature respectively in Figure 3.

Module	MOTA	MOTP	IDF	MT(%)	ML(%)	FP	FN	IDS
A+L	32.3	72.3	47.1	8.1	52.6	2,004	13,496	129
A+L+S	34.8	72.4	46.5	9.0	53.0	947	13,966	151
A+L+S+T	36.2	72.2	46.6	9.0	51.7	1799	13,079	94
A+L+S+Avg	33.1	72.2	37.1	6.4	54.7	888	14,386	176
A+L+S+Max	33.9	72.4	43.4	8.5	54.7	848	14,268	140

Table 2. Ablation study of the spatial temporal relation network.

Besides, we only keep the sequences that show up in the nearest $1.25F$ frames for enabling efficient inference.

4.3. Ablation Study

We follow [46] to split the 11 training sequences into train/val sets for ablation study.

Design of Feature Representation We first examine the effects of various design of feature representation in Table 1. All the experiments are based on the original appearance features without spatial-temporal reasoning.

The first three rows compare the effects of different appearance features *without the relation modules*. By only using unary appearance representation, it achieves 19.8 in terms of MOTA. By using cosine value alone, it gets 25.2 in MOTA. By using the hybrid features of both unary appearance and cosine value, the accuracy is significantly higher, reaching 29.8 in MOTA.

The last three rows compare the effects of different location features. By only utilizing the unary location features in Eqn. (10), 1.9 MOTA improvements is observed. By utilizing the motion features in Eqn. (11), 1.2 improvements in MOTA is observed. By combining both of them, we achieve 2.5 MOTA boosts. Also note that with the location features, the ID switch is significantly reduced, from 515 to 129.

The effects of Spatial-Temporal Relation Module Table 2 examines the effects of spatial-temporal relation module in improving the tracking accuracy. With relation reasoning along the spatial domain, the tracking accuracy improves by 2.5 in terms of MOTA. Significant reduction in FP is observed, indicating the topology encoded by spatial relation reasoning could help the association method to more accurately identify wrong associations. Further performing temporal relation reasoning, an additional 1.4 MOTA improvement is achieved. Note that our temporal relation reasoning is essentially a weighted average over all frame fea-

Table 3. Tracking Performance on 2DMOT2015 benchmark dataset.

Mode	Method	MOTA↑	MOTP↑	IDF↑	IDP↑	IDR↑	MT(%)↑	ML(%)↓	FP↓	FN↓	IDS↓	Frag↓	AR↓
Offline	MHT_DAM [25]	32.4	71.8	**45.3**	58.9	36.8	16.0	43.8	9,064	32,060	**435**	826	21.7
	NOMT [11]	33.7	71.9	44.6	**59.6**	35.6	12.2	44.0	**7,762**	32,547	442	**823**	**18.7**
	QuadMOT [51]	33.8	**73.4**	40.4	53.5	32.5	12.9	**36.9**	7,898	32,061	703	1,430	23.5
	JointMC [23]	**35.6**	71.9	45.1	54.4	**38.5**	**23.2**	39.3	10,580	**28,508**	457	969	19.3
Online	SCEA [20]	29.1	71.1	37.2	55.9	27.8	8.9	47.3	60,60	36,912	**604**	**1,182**	30.4
	MDP [58]	30.3	71.3	44.7	57.8	36.4	13.0	38.4	9,717	32,422	680	1,500	25.9
	CDA_DDAL [1]	32.8	70.7	38.8	58.2	29.1	9.7	42.2	4,983	35,690	614	1,583	24.2
	AMIR15 [46]	37.6	71.7	46.0	58.4	**38.0**	**15.8**	**25.8**	7,933	**29,397**	1,026	2,024	19.6
	ours	**38.1**	**72.1**	**46.6**	**63.9**	36.7	11.5	33.4	**5,451**	31,571	1,033	2,665	**16.1**

Table 4. Tracking Performance on MOT16 benchmark dataset.

Mode	Method	MOTA↑	MOTP↑	IDF↑	IDP↑	IDR↑	MT(%)↑	ML(%)↓	FP↓	FN↓	IDS↓	Frag↓	AR↓
Offline	NOMT [11]	46.4	76.6	**53.3**	73.2	**41.9**	18.3	41.4	9,753	87,565	**359**	**504**	18.6
	MCjoint [23]	47.1	76.3	52.3	**73.9**	40.4	**20.4**	46.9	6,703	89,368	370	598	19.8
	NLLMPa [30]	47.6	78.5	47.3	67.2	36.5	17.0	40.4	5,844	89,093	629	768	18.8
	FWT [18]	47.8	75.5	44.3	60.3	35	19.1	**38.2**	8,886	**85,487**	852	1,534	24.8
	GCRA [32]	48.2	77.5	48.6	69.1	37.4	12.9	41.1	**5,104**	88,586	821	1,117	21.9
	LMP [54]	**48.8**	**79.0**	51.3	71.1	40.1	18.2	40.1	6,654	86,245	481	595	**17.8**
Online	oICF [24]	43.2	74.3	49.3	73.3	37.2	11.3	48.5	6,651	96,515	**381**	**1,404**	31.8
	STAM [12]	46.0	74.9	50	71.5	38.5	14.6	43.6	6,895	91,117	473	1,422	29.3
	DMAN [63]	46.1	73.8	**54.8**	**77.2**	42.5	**17.4**	42.7	7,909	89,874	532	1,616	23.4
	AMIR [46]	47.2	**75.8**	46.3	68.9	34.8	14.0	41.6	**2,681**	92,856	774	1,675	22.9
	MOTDT [10]	47.6	74.8	50.9	69.2	40.3	15.2	38.3	9,253	85,431	792	1,858	23.5
	ours	**48.5**	73.7	53.9	72.8	**42.8**	17.0	**34.9**	9,038	**84,178**	747	2,919	**15.4**

Table 5. Tracking Performance on MOT17 benchmark dataset.

Mode	Method	MOTA↑	MOTP↑	IDF↑	IDP↑	IDR↑	MT(%)↑	ML(%)↓	FP↓	FN↓	IDS↓	Frag↓	AR↓
Offline	IOU [5]	45.5	76.9	39.4	56.4	30.3	15.7	40.5	**19,993**	281,643	5,988	7,404	36.5
	MHT_DLSTM [26]	47.5	**77.5**	51.9	71.4	40.8	18.2	41.7	25,981	268,042	2,069	3,124	28.8
	EDMT [8]	50.0	77.3	51.3	67	41.5	**21.6**	36.3	32,279	**247,297**	2,264	3,260	24.0
	MHT_DAM [25]	50.7	**77.5**	47.2	63.4	37.6	20.8	36.9	22,875	252,889	2,314	**2,865**	25.4
	jCC [22]	51.2	75.9	**54.5**	**72.2**	**43.8**	20.9	37	25,937	247,822	**1,802**	2,984	**20.3**
	FWT [18]	**51.3**	77	47.6	63.2	38.1	21.4	**35.2**	24,101	247,921	2,648	4,279	24.2
Online	PHD_GSDL [16]	48.0	**77.2**	49.6	68.4	39	17.1	**35.6**	23,199	265,954	3,998	8,886	32.5
	AM_ADM [49]	48.1	76.7	52.1	71.4	41	13.4	39.7	25,061	265,495	2,214	5,027	27.3
	DMAN [63]	48.2	75.9	55.7	**75.9**	44	19.3	38.3	26,218	263,608	2,194	5,378	26.6
	HAM_SADF [61]	48.3	**77.2**	51.1	71.2	39.9	17.1	41.7	**20,967**	269,038	**1,871**	**3,020**	25.2
	MOTDT [10]	**50.9**	76.6	52.7	70.4	42.1	17.5	35.7	24,069	250,768	2,474	5,317	23.1
	ours	**50.9**	75.6	**56.5**	74.5	**45.5**	**20.1**	37.0	27,532	**246,924**	2,593	9,622	**18.2**

tures. Hence we also compare it to some straight-forward aggregation methods, such as average summation and max-pooling along the frame dimension. These methods perform significantly worse than ours, proving the effectiveness of our temporal relation reasoning method.

4.4. Results on the MOT Benchmarks

We report the tracking accuracy on all of the three MOT benchmarks in Table 3, 4 and 5. We used the public detections for a fair comparison. Our method achieves the state-of-the-art tracking accuracy under online settings on all of the three benchmarks considering the major metrics of MOTA and AR (average rank).

5. Conclusion

This paper studies the object association problem for multi-object tracking (MOT). To build a robust similarity measure, we combine various cues, including appearance, location and topology cues through utilizing relation networks in spatial domains and further extending the relation networks to the temporal domain for aggregating information across time. The resulting approach is dubbed as spatial-temporal relation networks (STRN), which runs feed-forward and in end-to-end. It achieves the state-of-the-art accuracy over all online methods on all of the MOT15∼17 benchmarks using public detection.

References

[1] Seung Hwan Bae and Kuk-Jin Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(3):595–610, 2018. 8

[2] Seung-Hwan Bae and Kuk-Jin Yoon. 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*, pages 1218–1225, 2014. 2

[3] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, pages 4502–4510, 2016. 2, 3

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *J. Image Video Process.*, 2008:1:1–1:10, Jan. 2008. 6

[5] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy, August 29 - September 1, 2017*, pages 1–6, 2017. 2, 8

[6] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–6. IEEE, 2017. 6

[7] William Brendel, Mohamed Amer, and Sinisa Todorovic. Multiobject tracking as maximum weight independent set. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1273–1280. IEEE, 2011. 2

[8] Jiahui Chen, Hao Sheng, Yang Zhang, and Zhang Xiong. Enhancing detection model for multiple hypothesis tracking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017*, pages 2143–2152, 2017. 8

[9] Long Chen, Haizhou Ai, Chong Shang, Zijie Zhuang, and Bo Bai. Online multi-object tracking with convolutional neural networks. In *ICIP*, 2017. 2

[10] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *IEEE International Conference on Multimedia and Expo, ICME 2018*, pages 1–6, 2018. 8

[11] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *IEEE International Conference on Computer Vision, ICCV 2015*, pages 3029–3037, 2015. 8

[12] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *2017 IEEE International Conference on Computer Vision (ICCV).(Oct 2017)*, pages 4846–4855, 2017. 1, 2, 3, 7, 8

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 1

[14] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4099, 2015. 2

[15] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, Sept. 2010. 1, 6

[16] Zeyu Fu, Pengming Feng, Federico Angelini, Jonathon A. Chambers, and Syed Mohsen Naqvi. Particle PHD filter based multiple human tracking using online group-structured dictionary learning. *IEEE Access*, 6:14764–14778, 2018. 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 7

[18] Roberto Henschel, Laura Leal-Taixé, Daniel Cremers, and Bodo Rosenhahn. Improvements to frank-wolfe optimization for multi-detector multi-object tracking. *CoRR*, abs/1705.08314, 2017. 8

[19] Ju Hong Yoon, Chang-Ryeol Lee, Ming-Hsuan Yang, and Kuk-Jin Yoon. Online multi-object tracking via structural constraint event aggregation. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 1392–1400, 2016. 1, 2

[20] Ju Hong Yoon, Chang-Ryeol Lee, Ming-Hsuan Yang, and Kuk-Jin Yoon. Online multi-object tracking via structural constraint event aggregation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 8

[21] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. 2018. 2, 3, 4, 6, 7

[22] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation multiple object tracking by correlation co-clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 8

[23] Margret Keuper, Siyu Tang, Zhongjie Yu, Bjoern Andres, Thomas Brox, and Bernt Schiele. A multi-cut formulation for joint segmentation and tracking of multiple objects. *CoRR*, abs/1607.06317, 2016. 8

[24] Hilke Kieritz, Stefan Becker, Wolfgang Hübner, and Michael Arens. Online multi-person tracking using integral channel features. In *13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016*, pages 122–130, 2016. 8

[25] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. Multiple hypothesis tracking revisited. In *IEEE International Conference on Computer Vision, ICCV 2015*, pages 4696–4704, 2015. 8

[26] Chanho Kim, Fuxin Li, and James M Rehg. Multi-object tracking with neural gating using bilinear lstm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 200–215, 2018. 2, 8

[27] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2016. 1, 2

[28] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr. 2015. arXiv: 1504.01942. 1, 6

[29] 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. 1, 3

[30] Evgeny Levinkov, Jonas Uhrig, Siyu Tang, Mohamed Omran, Eldar Insafutdinov, Alexander Kirillov, Carsten Rother, Thomas Brox, Bernt Schiele, and Bjoern Andres. Joint graph decomposition & node labeling: Problem, algorithms, applications. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 1904–1912, 2017. 8

[31] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *In CVPR*, 2009. 7

[32] Cong Ma, Changshui Yang, Fan Yang, Yueqing Zhuang, Ziwei Zhang, Huizhu Jia, and Xiaodong Xie. Trajectory factory: Tracklet cleaving and re-connection by deep siamese bi-gru for multiple object tracking. In *IEEE International Conference on Multimedia and Expo, ICME 2018*, pages 1–6, 2018. 8

[33] Andrii Maksai, Xinchao Wang, François Fleuret, and Pascal Fua. Non-markovian globally consistent multi-object tracking. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2563–2573. IEEE, 2017. 1, 2

[34] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831. 1, 6

[35] Anton Milan, Seyed Hamid Rezatofighi, Anthony R Dick, Ian D Reid, and Konrad Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, volume 2, page 4, 2017. 1, 2

[36] Anton Milan, Stefan Roth, and Konrad Schindler. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):58–72, 2014. 2

[37] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957. 3

[38] Songhwai Oh, Stuart J. Russell, and Shankar Sastry. Markov chain monte carlo data association for multi-target tracking. *IEEE Trans. Automat. Contr.*, 54(3):481–497, 2009. 2

[39] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1201–1208. IEEE, 2011. 2

[40] Hamed Pirsiavash, Deva Ramanan, and Charless C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[41] Liangliang Ren, Jiwen Lu, Zifeng Wang, Qi Tian, and Jie Zhou. Collaborative deep reinforcement learning for multi-object tracking. In *Computer Vision - ECCV 2018 - 15th European Conference*, pages 605–621, 2018. 2

[42] 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, 6

[43] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *CoRR*, abs/1609.01775, 2016. 6

[44] A. Roshan Zamir, A. Dehghan, and M. Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 2

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 7

[46] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 300–311. IEEE, 2017. 1, 2, 3, 6, 7, 8

[47] Ricardo Sanchez-Matilla, Fabio Poiesi, and Andrea Cavallaro. Online multi-target tracking with strong and weak detections. In *ECCV 2016 Workshops*, pages 84–99, 2016. 1, 2

[48] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017. 2, 3

[49] Myung-Yun KIM Seong-Ho Lee and Seung-Hwan Bae. Learning discriminative appearance models for online multi-object tracking with appearance discriminability measures. *IEEE Access*, 2018. 8

[50] Guang Shu, Afshin Dehghan, Omar Oreifej, Emily Hand, and Mubarak Shah. Part-based multiple-person tracking with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1815–1821. IEEE, 2012. 2

[51] Jeany Son, Mooyeol Baek, Minsu Cho, and Bohyung Han. Multi-object tracking with quadruplet convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5620–5629, 2017. 2, 8

[52] Siyu Tang, Bjoern Andres, Miykhaylo Andriluka, and Bernt Schiele. Subgraph decomposition for multi-target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5033–5041, 2015. 2

[53] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Multi-person tracking by multicut and deep matching. In *European Conference on Computer Vision*, pages 100–111. Springer, 2016. 2

[54] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person reidentification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017. 8

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2, 3, 4, 6, 7

[56] Bing Wang, Li Wang, Bing Shuai, Zhen Zuo, Ting Liu, Kap Luk Chan, and Gang Wang. Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016*, pages 386–393, 2016. 1

[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. 2018. 2, 3

[58] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *2015 IEEE international conference on computer vision (ICCV)*, number EPFL-CONF-230283, pages 4705–4713. IEEE, 2015. 1, 2, 3, 7, 8

[59] Bo Yang and Ram Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1918–1925. IEEE, 2012. 2

[60] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 6

[61] Young-Chul Yoon, Abhijeet Boragule, Kwangjin Yoon, and Moongu Jeon. Online multi-object tracking with historical appearance matching and scene adaptive detection filtering. *CoRR*, abs/1805.10916, 2018. 8

[62] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[63] Ji Zhu, Hua Yang, Nian Liu, Minyoung Kim, Wenjun Zhang, and Ming-Hsuan Yang. Online multi-object tracking with dual matching attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018. 1, 2, 3, 7, 8