Online Multiple Object Tracking with Cross-Task Synergy

Song Guo*[1] Jingya Wang*[1,2] Xinchao Wang[3,4] Dacheng Tao[1]
[1]The University of Sydney, [2]ShanghaiTech University,
[3]National University of Singapore, [4]Stevens Institute of Technology
sguo2908@uni.sydney.edu.au wangjingya@shanghaitech.edu.cn
xinchao@nus.edu.sg dacheng.tao@sydney.edu.au

Abstract

Modern online multiple object tracking (MOT) methods usually focus on two directions to improve tracking performance. One is to predict new positions in an incoming frame based on tracking information from previous frames, and the other is to enhance data association by generating more discriminative identity embeddings. Some works combined both directions within one framework but handled them as two individual tasks, thus gaining little mutual benefits. In this paper, we propose a novel unified model with synergy between position prediction and embedding association. The two tasks are linked by temporal-aware target attention and distractor attention, as well as identity-aware memory aggregation model. Specifically, the attention modules can make the prediction focus more on targets and less on distractors, therefore more reliable embeddings can be extracted accordingly for association. On the other hand, such reliable embeddings can boost identity-awareness through memory aggregation, hence strengthen attention modules and suppress drifts. In this way, the synergy between position prediction and embedding association is achieved, which leads to strong robustness to occlusions. Extensive experiments demonstrate the superiority of our proposed model over a wide range of existing methods on MOTChallenge benchmarks. Our code and models are publicly available at `https://github.com/songguocode/TADAM`.

1. Introduction

The problem of multiple object tracking (MOT) has been studied for decades because of its broad applications such as robotics, surveillance, and autonomous driving. It aims to locate targets while maintain their identities to form trajectories across video frames. Recent research in the area of MOT mostly follows the paradigm of tracking-by-detection,

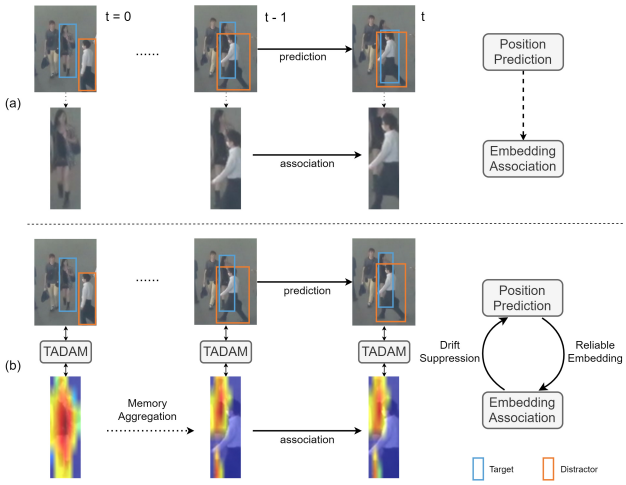

Figure 1. (a) Position prediction and embedding association in existing methods do not benefit each other during occlusion. The prediction of target position drifts and the extracted embeddings become noisy. (b) Our method brings synergy between the two tasks via proposed model to deal with occlusion.

which divides the MOT problem into two separate steps. Object detections are obtained independently in each frame first, then linked across frames through data association to form trajectories, where identity embeddings are usually adopted to distinguish objects during association. Such two-step procedure intuitively reveals two ways to improve the tracking performance. One is to augment detections, and the other is to enhance data association via embeddings.

Most existing online methods usually address only one of these two aspects for better tracking results, despite the fact that a common source of error, occlusions, affects both aspects. Unexpected occlusions often lead to miss detections due to overlapped objects, as well as increased difficulty for data association. Many online tracking approaches fill the gaps in detections during occlusions by predicting new positions of tracked targets, while a number of studies focus on generating more distinguishable embeddings

*Equal contributions

to be associated throughout occlusions. Although a few recent works attempted to tackle both simultaneously, position prediction and embedding association were treated as two individual tasks. How to make them benefit each other has not been well explored.

Common prediction methods seldom take interaction between objects into account, thus position prediction itself is not strong enough when dealing with occlusions. Making predictions under heavier occlusions often leads to drifted bounding boxes, where the predicted position of a target starts to follow a neighbouring object. The embedding extracted for association then deteriorates due to the wrongly predicted bounding box. This may lead to association errors that propagate over successive frames. Making predictions harm associating embeddings instead of helping in such cases. Meanwhile, improving embeddings alone only reduces errors at association stage, which does not help preventing position prediction errors at first hand. As such, there is no real synergy between the two tasks by treating them as two separate problems, as demonstrated in Fig. 1.

In this paper, we propose a unified model where position prediction and embedding association are jointly optimized with mutual benefits, boosting tracking performance by enhancing robustness to occlusions. To bring a real synergy, we make one task participate in the other's process. The two tasks are bridged by a link consisting of a target attention module and a distractor attention module, as well as a discriminative memory aggregation. Identity embeddings optimized for association are not only used in calculating affinity, but also applied to generate focus on a target as well as suppress drift through attention modules. In this way, the position prediction is equipped with identity-awareness and becomes sensitive to nearby objects, where more correct predictions can be performed under heavier occlusions without drifts. With better predictions during occlusions and attentions on the target, higher-quality embeddings can be extracted. Such more reliable embeddings then participate in the attention generation for better focus on the target. As a result, position prediction and embedding association are involved with each other, thus form a positive feedback loop with mutual benefits. The synergy is further amplified by an identity-aware memory aggregation, as richer holistic embeddings accumulated over time enable more robust attention generation. Consequently, tracking performance in complicated scenes with occlusions is boosted. We jointly optimize the position prediction, embedding association, and all proposed modules under a unified end-to-end model. To the best of our knowledge, we are the first to achieve synergistic joint optimization on the two tasks.

The main contributions of this paper are listed as follows:

- We propose a unified online MOT model that brings mutual benefits between position prediction and embedding association, thus achieving stronger robustness to occlusions.
- We apply temporal-aware target attention and distractor attention as well as an identity-aware memory aggregation to link the two tasks.
- Our tracker achieves state-of-the-art performance on MOTChallenge benchmarks with public detections.

2. Related Work

Challenges in tracking-by-detection. The tracking-by-detection paradigm has been commonly adopted in most modern online multiple object trackers [5, 52, 53, 26, 33, 32, 54, 49, 10, 63, 56, 3, 57, 30]. Off-the-shelf detectors like DPM [38], Faster R-CNN [39], SDP [58], YOLO [38], or SSD [31] are first applied to discover objects in each incoming frame, then followed by data associations, where objects found in different frames are linked to form trajectories. Although earlier approaches like JPDA [20, 41], MCMC-DA [36], and MHT [6, 60, 25] evaluate all possible associations and directly form most probable trajectories in one step, they have been considered inefficient and not scalable in modern online complex MOT scenes. Methods adopting the tracking-by-detection paradigm face challenges in its two steps when tracking complicated scenarios with more occlusions. On the one hand, detections given by the detector become inaccurate or even missing due to occlusions. Such imperfection often gives rise to intermittent or fragmented tracklets and therefore degrades the tracking result. On the other hand, associating objects under complex scenes requires association measurement with stronger discriminability among objects with different identities. To this end, many online MOT methods aim to improve MOT performance by tackling either of these two issues.

Position prediction with visual cue. To fill gaps in detections, many works propose to infer locations of objects when they are not correctly given. In offline methods where all frames are provided and processed together, interpolations are performed to deduce intermediate positions once two object instances across multiple frames are confirmed to have the same identity [45, 8, 37, 24, 51, 27, 40, 22, 46]. However, such batch processing is not applicable in online methods where decisions must be made without access to data beyond the latest frame. As a result, online methods adopt prediction on target positions to deal with the gaps. Prediction can be made solely with motion models, such as a linear model like the Kalman Filter applied in [5, 49] and non-linear model like LSTM used in [35, 1, 19, 61], but relying on motion only cannot achieve comparable performance with approaches utilizing visual cues for position prediction. For example, correlation filters have been ap-

plied to estimate new positions by finding the highest response in a new frame with visual features extracted from previous frames [49, 16, 62]. Single object tracking (SOT) trackers like ECO [15], SiamFC [4], and SiamRPN [28] can be adopted in MOT by initiating one tracker for each target [63, 18, 13]. While they do eliminate some gaps in detections, such trackers lack the ability to differentiate objects of the same class, and are therefore vulnerable to occlusions by distractors. As frequency of occlusion is much higher and distractors are less distinguishable in MOT, special design is usually necessary to make SOT trackers to fit MOT framework. Bounding box regression in the second step of two-stage detectors like Faster R-CNN [39] can be used as a predictor for new positions, by extracting features with previous bounding boxes passed from previous frame and infer displacements of boxes [3]. However, its power is still limited when dealing with heavier occlusions. To address the issue of occlusion in online position prediction, we propose to enforce stronger focus on targets and strengthen resistance to distractors.

Association with identity embeddings. Building a more reliable data association measurement is another direction to improve MOT performance. Earlier works link detections across frames with bounding box Intersection-over-Union (IoU) [5, 7, 47, 50], which is fast but not often inaccurate. Extracting an appearance embedding from each bounding box can establish a more discriminative association metric to distinguish objects of different identities. The identity embeddings can be used as main source of association [54, 10], or in conjunction with other features like motion feature [44, 49, 43]. Such methods usually need a dedicated model trained with extra datasets, which incurs non-trivial cost in computation. More sophisticated association metrics can be built on fusion of identity embeddings with motion features [55, 13, 59], or with fine-grained visual feature like body joints for pedestrian tracking [48, 23]. Other approaches like layered tracking [14, 2, 9] may also help identifying different targets. Accordingly, they all introduce much higher costs in the procedures of model designing and training.

More recently, UMA [59] proposed to integrate embedding generation into its position prediction with a triplet structure, while DeepMOT [57] adopted an embedding head to produce identity embeddings simultaneously with regression-based position prediction. Such multi-task design lowers the cost for training association metric, but position prediction and association are treated as two individual tasks and their outcomes are not benefiting each other. While UMA [59] designed a task-specific feature transformation to make the two tasks compatible under SOT framework, training an integrated embedding head in DeepMOT [57] has no impact on prediction results compared with having an externally trained embedding model. We show that the two tasks can work together in a more synergistic way that one of them participate in the improvement of another in tracking, which is essentially useful in complex scenes with occlusions.

3. Proposed Method

In this work, we propose a unified model that brings mutual benefit between position prediction and data association, so that robustness to occlusions is enhanced and tracking performance is boosted. To achieve this, we introduce temporal-aware target attention and distractor attention to form better focus on targets and suppress interference from distractors, as well as an identity-aware memory aggregation scheme for more robust attention generations. We name it by designed components as TADAM, where TA and DA refers to the target attention and the distractor attention, while M denotes the memory aggregation. All components are trained with the same data source within a unified model. The overall framework of the proposed method is illustrated in Fig. 2.

Problem formulation. A tracked target object formed in a given sequence before frame t can be denoted as T^{ta}_{t-1}, where its bounding box in frame $t-1$ is described by B^{ta}_{t-1}. A nearby distractor is described by T^{di}_{t-1} with its bounding box B^{di}_{t-1}. F^{ta}_t represents the target's new position prediction feature extracted at frame t using B^{ta}_{t-1}. While E^{ta}_t and E^{di}_t stands for similarly extracted latest identity embeddings of the target and its distractor respectively, their historical embedding references are given by E^{ta}_r and E^{di}_r.

3.1. Preliminary of position prediction by regression

We adopt a regression-based position regression tracker [3] as a baseline, since it outperforms other prediction method with visual cues. It trains a two-stage Faster R-CNN detector with provided data, where an RPN is trained to generate coarse proposals boxes in the first stage, and a regression head together with a classification head are trained to refine boxes and deduce classes of objects inside boxes. During tracking, the first RPN stage is discarded, while the trained regression head is exploited to predict new position of a tracked target B^{ta}_t from prediction feature F^{ta}_t extracted at its previous location B^{ta}_{t-1}, with the classification head giving the confidence for the prediction. Fig. 2 (a), without connections with (b), illustrates the tracking procedure of this position prediction by regression. Embedding E^{ta}_t used for association is then obtained with an embedding extraction process. The power of position prediction mainly comes from inferring a tightly fitted bounding box with a given less accurate box, and it is trained with smooth L1 loss on displacements at four sides as adopted in [39].

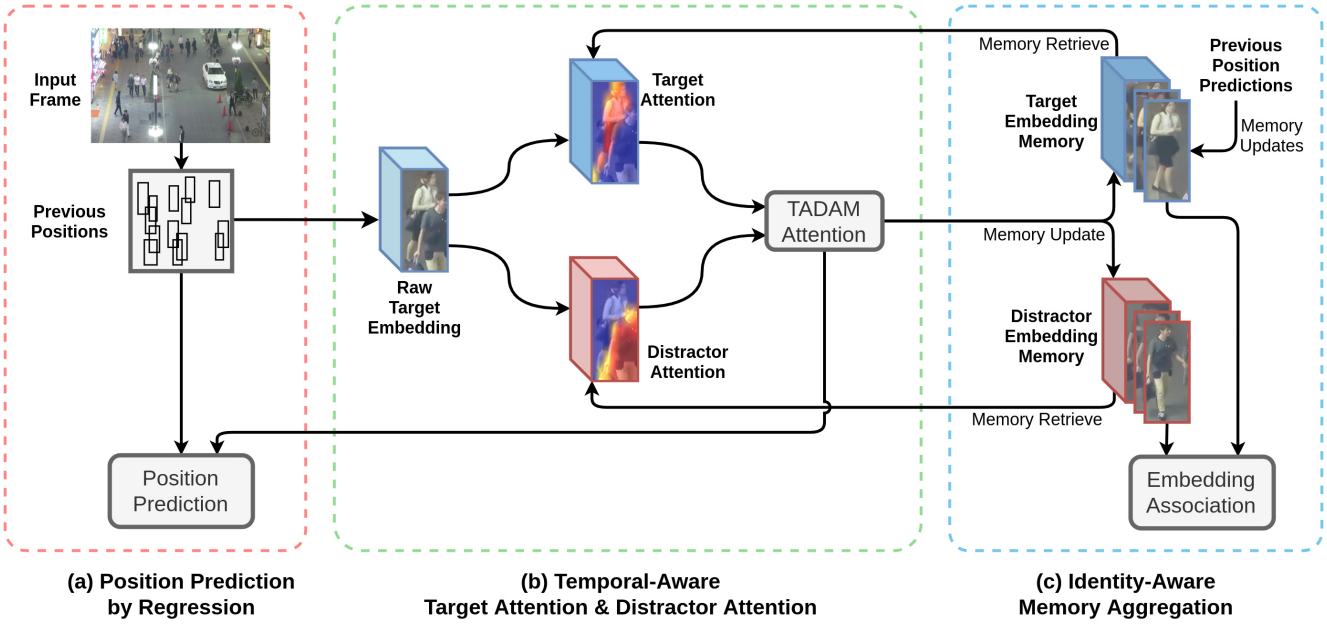

Figure 2. Model structure. The two tasks of position prediction and embedding association are bidirectionally linked by proposed modules to form synergy.

Meanwhile, the classification head is learned with a cross-entropy loss between inferred classes and ground truth class annotations of input bounding boxes. Such position prediction approach gets rid of data association for targets being actively tracked, while matching through Hungarian algorithm is still necessary to search for potential reappearance of lost targets among new detections by comparing identity embeddings. Our proposed method aims to bring cross-task synergy on top of the position prediction design.

3.2. Temporal-Aware Target Attention and Distractor Attention

When a position prediction is performed on a target T^{ta} from frame $t-1$ to t, the new prediction feature F_t^{ta} in frame t is extracted with its previous bounding box B_{t-1}^{ta} and its new position can be predicted with F_t^{ta}. However, when a distractor T^{di} is nearby and its bounding box, B_{t-1}^{di} has a large overlap with B_{t-1}^{ta}, making a correct prediction becomes difficult. Suppose T^{ta} is occluded by T^{di}, namely T^{di} is in front, then the predicted new bounding box of T^{ta} will tend to be closer to T^{di}, as F_t^{ta} contains a large portion from F_t^{di} that actually belongs to T^{di}. Continuing position prediction in such scenario will lead to a gradual drift of B_t^{ta} onto B_t^{di}. To overcome such dominance, we introduce a target-attention (TA) module to augment regions in F_t^{ta} that belong to T^{ta} for better focus, as well as a distractor-attention (DA) module to suppress parts in F_t^{ta} belonging to T^{di} to reduce interference. The target attention is computed between the target's latest raw identity embedding E_t^{ta} and its historical aggregated embedding reference E_r^{ta}, while the distractor attention is generated with E_t^{di} and the distractor's reference E_r^{di}. For simplicity, the distractor of a tracked target is selected as another nearby tracked target with largest IoU, where the one with highest overlapping is picked in case of multiple distractors. With the two attentions applied on F_t^{ta} to obtain a refined prediction feature $\tilde{F}_t^{ta}$, a better position prediction of B_t^{ta} can be performed. To further enhance the robustness of attention modules, a discriminative memory aggregation is designed to provide aggregated references of objects over time to make the attention modules discriminative and temporal-aware, which is covered in Sec. 3.3.

Discriminative aggregated non-local attention. To enhance or suppress regions in prediction feature F_t^{ta} for better prediction, we compute an attention projected from a reference embedding E_r^{ta} to the newly extracted raw embedding E_t^{ta}. The dimension of embeddings is $\mathcal{R}^{C\times H\times W}$, where C stands for channel, H for height, and W for width. $E_{t_i}^{ta}$ and $E_{r_j}^{ta}$ represents two points of dimension $\mathcal{R}^C$ on E_t^{ta} and E_r^{ta} respectively, where $i, j \in [1, HW]$ denotes two arbitrary spatial locations. With a discriminative aggregated target reference embedding $E_{r_j}^{ta}$ as input, where the aggregation process is introduced in Sec. 3.3, an aggregated non-local target attention from historical memory reference $E_{r_j}^{ta}$ onto $E_{t_i}^{ta}$ can then be described as follows.

$$f(E_{t_i}^{ta}, E_{r_j}^{ta}) = \theta(E_{t_i}^{ta})\phi(E_{r_j}^{ta})\rho(E_{r_j}^{ta}), \quad (1)$$

where θ and ϕ are convolution layers for computing correlation between the two, while ρ is another convolution layer to generate an representation of $E^{ta}_{r_j}$ for output.

By serializing all location pairs j on $E^{ta}_{r_j}$ and i on $E^{ta}_{t_i}$, we obtain an overall non-local attention from the reference E^{ta}_r onto the new embedding E^{ta}_t. Since E^{ta}_r is an aggregation with identity-aware memory, this process becomes a discriminative aggregated non-local attention between the target's historical references and its raw embedding obtained in a new frame.

Similarly, the discriminative aggregated non-local distractor attention from a distractor reference embedding $E^{di}_{r_j}$ onto $E^{ta}_{t_i}$ is given by equation below.

$$g(E^{ta}_{t_i}, E^{di}_{r_j}) = \theta(E^{ta}_{t_i})\phi(E^{di}_{r_j})\rho(E^{di}_{r_j}), \tag{2}$$

While locations in the computed target attention with larger values indicate that those parts are more likely belong to the target, regions with higher response in distractor attention imply their greater probabilities of being parts of the distractor. We can then enhance the prediction with computation of refined prediction feature $\tilde{F}^{ta}_t$ given as follows.

$$\tilde{F}^{ta}_t = F^{ta}_t \oplus w[f(E^{ta}_t, E^{ta}_r) \ominus g(E^{ta}_t, E^{di}_r)], \tag{3}$$

where f and g stand for vectorized version of operations in Eq. 1 and Eq. 2, while $\oplus$ and $\ominus$ denotes element-wise addition and subtraction. w denotes a weight to regulate attention output amount. E^{ta}_t is the newly extracted raw embedding in frame t, while E^{ta}_r and E^{di}_r are the discriminative target reference and distractor reference retrieved from respective memories.

The combined temporal-aware target attention operation and distractor attention computation form TADAM attention for our model, as depicted in Fig. 2 (b). With $\tilde{F}^{ta}_t$ used for position prediction as mentioned in Sec. 3.1, the target's new position B^{ta}_t can then be predicted with more focus on itself and less interference from its distractor. Consequently, more correct predictions can be made under heavier occlusions, therefore allow collecting less noisy identity embeddings for data association and memory aggregation, which is discussed in Sec. 3.3.

Adaptive weight in target and distractor attention. While enhancement from TA and suppression from DA make position prediction more focused and less distracted, they are not necessarily useful for easy cases, especially on targets with little to no occlusion. Furthermore, applying a fixed weight w as in Eq. 3 for TA output and DA output for all degrees of occlusions is suboptimal. Targets undergoing heavier occlusions should expect larger enhancement and stronger suppression, while those with little overlaps with neighbours should be left as is with no processing. To address this issue, an adaptive weight for target attention and distractor attention is designed as follows.

$$w = \frac{max(iou(B^{ta}_t, B^{di}_t) - o_{min}, 0)}{1 - o_{min}}, \tag{4}$$

where $iou(,)$ stands for computation of the IoU between two input boxes, B^{ta}_t represents a target's box and B^{di}_t refers to the box of its distractor. o_{min} gives the minimum level of overlapping for the weight to take effect, and $max(,)$ outputs larger value of the two inputs. This value is computed per target-distractor pair.

The weight becomes non-zero and ranges between 0 to 1 when the computed IoU between a target and its distractor exceeds o_{min}, otherwise assigned to 0. Output from target attention and distractor attention are adaptively regulated by this weight before participating further refinement. Namely, target attention and distractor attention do not engage in easy cases with insignificant occlusions, while larger portion of their outcomes are used to handle more severe occlusions.

3.3. Identity-Aware Memory Aggregation

Reference embeddings of targets and those of distractors that participate in the target attention module and the distractor attention module respectively are described in Sec. 3.2. Although storing embeddings formed in the previous frame is an intuitive and accessible way to obtain references, embeddings obtained in such approach are usually noisy in more complex scenarios with heavier occlusions. To enhance the robustness of attention computations, we propose an identity-aware memory aggregation to accumulate more holistic references as inputs to the TA and DA attention modules, so that position predictions and embeddings can both be further boosted.

In each frame, target attention and distractor attention computed with newly obtained raw embedding E^{ta}_t of a target is used to produce refined prediction feature $\tilde{F}^{ta}_t$. Similarly, E^{ta}_t itself is also processed by the attentions to form refined embedding $\tilde{E}^{ta}_t$ for association and aggregation, like the generation of $\tilde{F}^{ta}_t$ in Eq. 3. As the dimension of $\tilde{E}^{ta}_t$ to be aggregated is $\mathcal{R}^{C\times H\times W}$, and the reference feature produced after aggregation needs spatial dimensions to be used for attention computation, we have to keep spatial information before and after aggregation. In addition, for a holistic aggregation, we expect the aggregation be able to automatically determine whether an input is worth updating, rather than naive accumulation that stores every input. To address these concerns, we design a discriminative memory module with convolutional gated recurrent unit (GRU), where matrix multiplications in GRU are replaced with convolutions. This builds a memory for temporal relation across frames as well as keep spatial dimensions. The update of aggregated embedding memory is described as follows.

$$E_{r_t} = update(\tilde{E}_t, E_{r_{t-1}}), \tag{5}$$

where $update()$ is the memory update function, $E_{r_{t-1}}$ stands for previous state of the memory, and E_{r_t} refers to aggregated embedding updated with $\tilde{E}_t$ and $E_{r_{t-1}}$. Note that this update process applies to both targets and distractors, where E is replaced by E^{ta} for target embedding aggregation and substituted with E^{di} for distractor embedding aggregation, same as the notations used in Eq. 3. The $update()$ follows the GRU state calculation in [11], while dot products are replaced by convolutional layers to allow two-dimensional inputs.

Joint learning of memory aggregation and embedding extraction. Target reference embedding and distractor reference embedding retrieved from respective memories are required to be well separated in their embedding space. Otherwise, outputs of attention modules would not produce correct attentions on designated locations, but generates similar level of responses across regions belonging to any object instead. Meanwhile, since embeddings used for data association are originally required to be distinguishable without aggregation, we expect all embeddings to be discriminative to identities both before and after memory aggregation. To this end, we optimize the temporal aggregation and embedding formation with a joint discriminative learning process. An embedding of a target extracted from backbone network output firstly goes through four state initialization convolution layers, which make it distinguishable among different identities. If it is not the first embedding in a sequence, then we update it in the way described in Eq. 5 for further aggregation and learn discriminability as well. Both processes are jointly trained with two discriminative identity losses, one cross-entropy loss computed between predicted identities and ground truth identities, and another triplet loss to maximize inter-identity difference and minimize intra-identity distance.

When we feed an input of length 1 for aggregation, we obtain a resultant embedding without actual aggregation. As such, we can use the memory module as an embedding extraction approach in our framework by feeding single-length inputs. With longer input of same target, the memory aggregation continuously generates aggregated discriminative embeddings. In this way, we achieve joint optimization of memory aggregation and embedding extraction.

Synergy between tasks. With aggregated discriminative embeddings, more correct attentions can be obtained in attention modules with focus on targets and suppression on distractions. Applying the attentions in position prediction lead to stronger resistance to drifts. Conversely, with more correct and prolonged predicted bounding boxes in complex scenes, we can accumulate more embeddings which are refined by attention modules, and feed them for more reliable representation through the aggregation. In this regard, temporal-aware target attention and distractor attention described in Sec. 3.2 and identity-aware temporal aggregation presented in Sec. 3.3 closely work together to form a link between the task of position prediction and the task of embedding association, thereby bring a cross-task synergy.

4. Experiments

The tracking performance of our proposed method is evaluated on MOTChallenge benchmarks of MOT16, MOT17, and MOT20. We also conduct an ablation study with analyses to verify the effectiveness of our design.

MOTChallenge and metrics. The MOTChallenge provides benchmarks for comparing performance of different multi-object tracking algorithms. It contains multiple pedestrian tracking scenes with various conditions like lighting, crowdness, and camera motion. The most commonly accepted challenges for benchmark are MOT16 and MOT17, both consisting of 7 video sequences for training and 7 sequences for testing. All sequences are provided with public detections. MOT17 comes with detections from three different object detectors, DPM [17], Faster R-CNN [39], and SDP [58]. MOT16 contains the same sequences, but only has DPM as public detection source and its ground truth boxes for training are less accurate than MOT17. A newer benchmark MOT20 aims to test performance under extremely crowded scenarios, which contains 4 sequences for training and 4 sequences for testing. Compared to MOT16 and MOT17, MOT20 is not yet tested by many methods due to its late emergence, but it can still provide insights on directions in complex scenes. Performance of a tracker is evaluated from several aspects by a number of metrics, while the main factors are Multiple Object Tracking Accuracy (MOTA) and ID F1 score (IDF1). MOTA measures overall performance of a tracker by evaluating errors from three sources, namely False Negatives (FN), False Positives (FP) and Identity Switches (IDS) [34]. IDF1 focuses on the quality of assigned identities on detections with a uniformed scale [42]. To make fair comparisons among trackers, all tracking experiments are based on the public detections provided by the MOTChallenge. Particularly, object trajectories are only initiated after the first time they appear in provided public detections.

Implementation details Experiments are conducted on a desktop with RTX 2080 Ti GPU using PyTorch. We pretrain our backbone of ResNet101 [21] parameters on COCO dataset [29], then train on respective MOT dataset with all ground truth labeled objects with a minimum visibility of 0.1. The RPN anchor ratios are set to $\{1.0, 2.0, 3.0\}$. We sample 2 image frames per batch and pick 256 proposals in each image from all RPN proposals with a positive proposal

	Method	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓
MOT16	MOTDT16[10]	47.6	50.9	9253	85431	792
	KCF16[12]	48.8	47.2	5875	86567	906
	DeepMOT[57]	54.8	53.4	2955	78765	645
	Tracktor++V2[3]	56.2	54.9	**2394**	76844	617
	GSM[30]	57.0	58.2	4332	73573	**475**
	TADAM(ours)	**59.1**	**59.5**	2540	**71542**	529
MOT17	MOTDT17[10]	50.9	52.7	24069	250768	2474
	FAMNET[13]	52.0	48.7	14138	253616	3072
	UMA[59]	53.1	54.4	22893	239534	2251
	DeepMOT[57]	53.7	53.8	11731	247447	1947
	Tracktor++V2[3]	56.3	55.1	**8866**	235449	1987
	GSM[30]	56.4	57.8	14379	230174	**1485**
	TADAM(ours)	**59.7**	**58.7**	9676	**216029**	1930
MOT20	SORT20[5]	42.7	45.1	27521	264694	4470
	Tracktor++V2[3]	52.6	**52.7**	**6930**	236680	**1648**
	TADAM(ours)	**56.6**	51.6	39407	**182520**	2690

Table 1. Comparison with modern online methods on provided public detections of MOTChallenge benchmarks. Best result in each metric is marked in **bold**.

sampling ratio of 0.75. We warm up the training with memory module for embedding and aggregation learning for 3 epochs to achieve faster convergence, where learning rate is set to 0.2. We then jointly train all components for 12 epochs with an initial learning rate of 0.002 that decays by 0.5 in every 3 epochs. o_{min} is set to 0.2 empirically. It takes around 9 hours for training on MOT16 and MOT17 to finish on a single GPU, and 15 hours on MOT20 with two GPUs.

4.1. Benchmark Evaluation

The performance of our tracker is evaluated on the test set of MOTChallenge benchmarks. For MOT17, the final result is computed on all three provided public detection sets. We compare against modern online method that are officially published on the benchmark with peer-reviews.

As shown in Table 1, the benchmark results show the superior performance of our method over other published online public methods on MOT16, MOT17, and MOT20. It is noteworthy that we achieve best IDF1 and FN, in addition to the highest MOTA, on MOT16 and MOT17. Since we employ two attention modules to assist position prediction, one for enhancing focus on targets themselves and another for reducing distractions from neighbors, identities of objects are well taken care of, therefore it is not surprising to see that our proposed method performs well in terms of identity correctness indicated by IDF1. Although IDS of our method is not the best, ranking second on MOT16 and third on MOT17, it performs sufficiently well considering the minimal design of our embedding extraction. Methods with better IDS adopt more complicated models, like GSM [30] trained a graph similarity model. Compared with DeepMOT [57] that also employed an integrated identity embedding optimization, we have lower IDS on both MOT16 and MOT17 with much higher IDF1 on both datasets. On the other hand, our tracker has more FP than Tracktor++V2 [3], which indicates our enhanced position prediction is not always correct and could have produced FP in the process. Still, decrease in FN exceeds rise in FP by far compared with their result in both MOT16 and MOT17, which implies the effectiveness of our design. For MOT20, our method presents a best result among published online methods, where FN is notably less than other methods. To sum up, the benchmark results demonstrate our tracker's strong performance, and detailed analysis of it is conducted through an ablation study in Sec. 4.2.

Setup	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓
w/o TA & DA	65.9	71.1	597	37501	208
w/o DA	66.4	71.2	**462**	37060	191
w/o TA	66.7	71.3	473	36748	**188**
w/o adaptive weight	66.0	68.5	679	37322	242
w/o memory aggregation	66.5	67.5	552	36848	232
Full model	**67.0**	**71.6**	583	**36287**	197

Table 2. Ablation study on components on FRCNN of MOT17 train set. "TA" stands for target-awareness, "DA" for distractor-awareness. Best result in each metric is marked in **bold**

4.2. Ablation Study

An ablation study is conducted on MOT17 training set with provided FRCNN public detections. As shown in Table 2, we remove proposed components to see their contributions to our method. With both TA and DA modules removed, the tracking performance measured by MOTA decreases by 1.1, with worse result in all other metrics. The difference in performance mainly comes from the number of FN, where FN in full model significantly reduces, while FP also slightly decreases. This shows that more correct predictions are made with TA and DA enabled. Meanwhile, better IDF1 and fewer IDS indicate that the full model performs better in distinguishing identities with TA and DA. We see better prediction results with attention modules, which shows the benefit of adopting attentions from memory aggregated embeddings in prediction. Meanwhile, embeddings used for association are also improved, as demonstrated by the stronger discriminability. The higher performance in both tasks confirms the synergy in between.

Compared with the case where no TA and DA is employed, introducing TA without DA improves MOTA by 0.5, while applying DA alone leads to 0.8 higher MOTA. Specifically, least FP is seen in TA only setup, while lowest IDS is observed with only DA enabled. They both can improve tracking performance on their own and produce attention to bring synergy. However, the improvement is not as large as having them work together.

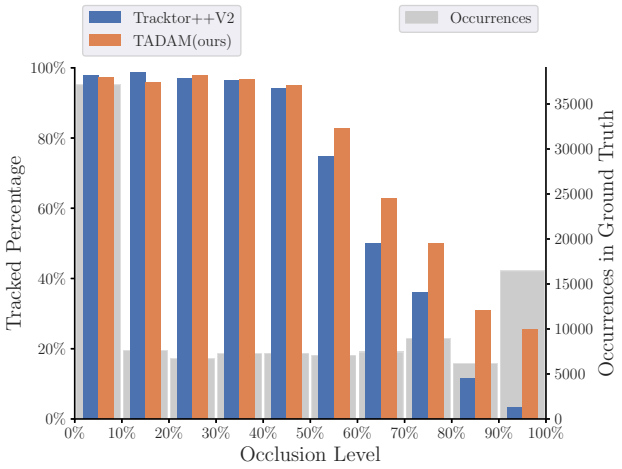

Figure 3. Occlusion level vs tracked percentage. Wider gray bars show the occurrence of ground truth object bounding boxes in each occlusion level interval, while narrower colored bars illustrate the percentage of objects tracked within each interval for their respective method. Note that occurrences and tracked percentages are not drawn in same unit.

Effect of adaptive weight. To verify the necessity of adaptive weight, we treat all cases in the same way by removing adaptive weight regardless of occlusion levels. A drop of 1.0 in MOTA is observed comparing with the full model, which makes the performance even worse than running with TA or DA alone, though still slightly higher than that without both TA and DA. This implies that naively applying TA and DA for all cases is indeed a suboptimal approach. While occlusions have to be taken care of, it can be seen from Fig. 3 that easy cases with slight occlusions dominates the dataset, where prediction without attention modules works sufficiently well. Therefore, it is necessary to leave them untouched and apply stronger attention only for harder cases with severer occlusions.

Effect of memory aggregation. We also conduct an experiment to see the benefit of our memory aggregation. Instead of using memory aggregation, we only store target and distractor embeddings extracted in previous frame as references for TA and DA. Without the aggregation, we observe a decrease in MOTA of 0.5, as well as worse result in FP, FN, IDF1, and IDS. This indicates that the discriminative memory aggregation significantly helps the TA and DA to form robust attentions for both prediction and embedding, therefore leads to stronger performance in tracking.

Performance in different occlusion levels. An intuitive way to verify the robustness of a tracker to occlusions is to check how many of the occluded targets are being tracked under different occlusion levels. In the annotation of MOT17 train set, we have access to the levels of occlusion of all ground truth objects. With occlusion levels quantized into intervals of 10%, occurrence of occlusion degrees within each interval can be counted to show the distribution of object visibility. Meanwhile, by calculating the percentage of ground truth objects that are tracked in each interval, we can evaluate a tracker's performance under different occlusion levels. To determine if a ground truth bounding box is covered by a tracker, its IoU is computed with all tracked targets' boxes in the same frame and compared with a threshold of 0.5 [42]. We compare our tracking result with Tracktor++V2 [3] as shown Fig. 3.

It is observed that our tracker has very similar tracked percentage with Tracktor++V2 [3] when object occlusion level is less than 50%. This implies that such position prediction with visual cues has achieved solid performance under low to medium levels of occlusion and leave little room for improvement. For objects with more than 50% occluded, our framework shows its advantage. The higher the occlusion level, the larger the performance boost with our tracker. This is highlighted on occlusion level $>$90%, where our tracker achieves around 25% tracked ratio against approximately 5%. This experiment confirms that our tracker does have better performance when dealing with occlusions. Nevertheless, how to track better with extremely low visibility could still be a direction in future research.

5. Conclusion

In this paper, we have proposed a method that jointly optimizes position prediction and embedding association with mutual benefits. The two tasks are bridged by a target attention module and a distractor attention module, as well as an identity-aware memory aggregation. The designed attention modules strengthen prediction by forcing more attention on targets and less interference from distractors, which enables extraction from more reliable embeddings for the association. On the other hand, these embeddings are exploited to form attention in prediction with the help of the memory aggregation module, and therefore assist in suppressing drifts. In this way, a synergy between the two tasks has been formed, which shows strong robustness in complex scenarios with heavy occlusions. In our experiments, we have demonstrated the remarkable performance of our method and the effectiveness of proposed components with extensive analyses. We expect that our method can pave the way for future research to reveal potential cross-task benefits in multi-task problems like MOT.

Acknowledgement

This work was supported in part by Australian Research Council Projects under Grant FL-170100117, Grant DP-180103424 and Grant IH-180100002, Grant IC-190100031.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *CVPR*, 2016.

[2] Jan Bandouch and Michael Beetz. Tracking Humans Interacting with the Environment Using Efficient Hierarchical Sampling and Layered Observation Models. In *ICCV Workshop*, 2009.

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019.

[4] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip H S Torr. Fully-Convolutional Siamese Networks for Object Tracking. Technical report.

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *ICIP*, 2016.

[6] Samuel S. Blackman. Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine*, 19(1 II):5–18, 2004.

[7] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-Speed Tracking-by-Detection without Using Image Information. In *AVSS*, 2017.

[8] Guillem Brasó and Laura Leal-Taixé. Learning a Neural Solver for Multiple Object Tracking. In *CVPR*, 2020.

[9] Jason Chang and John W Fisher III. Topology-Constrained Layered Tracking with Latent Flow. In *ICCV*, 2013.

[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 *ICME*, 2018.

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*, 2014.

[12] Peng Chu, Heng Fan, Chiu C. Tan, and Haibin Ling. Online Multi-object Tracking with Instance-aware Tracker and Dynamic Model Refreshment. In *WACV*, 2019.

[13] Peng Chu and Haibin Ling. FAMNet: Joint Learning of Feature, Affinity and Multi-dimensional Assignment for Online Multiple Object Tracking. In *ICCV*, 2019.

[14] Congxia Dai, Yunfei Zheng, and Xin Li. Layered Representation for Pedestrian Detection and Tracking in Infrared Imagery. In *CVPR Workshop*, 2006.

[15] Martin Danelljan, Goutam Bhat, Shahbaz Khan, and Michael Felsberg. ECO: Efficient Convolution Operators for Tracking. In *CVPR*, 2017.

[16] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to Track and Track to Detect. In *ICCV*, 2017.

[17] Pedro F Felzenszwalb, Ross B Girshick, David Mcallester, and Deva Ramanan. Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, sep 2010.

[18] Weitao Feng, Zhihao Hu, Wei Wu, Junjie Yan, and Wanli Ouyang. Multi-Object Tracking with Multiple Cues and Switcher-Aware Classification. Technical report, 2019.

[19] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Tracking by Prediction: A Deep Generative Model for Mutli-Person localisation and Tracking. In *WACV*, 2018.

[20] Thomas E. Fortmann, Yaakov Bar-Shalom, and Molly Scheffe. Multi-target tracking using joint probabilistic data association. In *19th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes*, 1980.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[22] Z. He, X. Li, X. You, D. Tao, and Y. Y. Tang. Connected component model for multi-object tracking. *IEEE Transactions on Image Processing*, 25(8):3698–3711, 2016.

[23] Roberto Henschel, Yunzhe Zou, and Bodo Rosenhahn. Multiple People Tracking using Body and Joint Detections. In *CVPR Workshop*, 2019.

[24] Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, and Paul Swoboda. Lifted Disjoint Paths with Application in Multiple Object Tracking. In *ICML*, 2020.

[25] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple Hypothesis Tracking Revisited. In *ICCV*, 2015.

[26] Long Lan, Xinchao Wang, Gang Hua, Thomas S. Huang, and Dacheng Tao. Semi-online multi-people tracking by re-identification. *IJCV*, 128:1937–1955, 2020.

[27] Long Lan, Xinchao Wang, Shiliang Zhang, Dacheng Tao, Wen Gao, and Thomas S. Huang. Interacting tracklets for multi-object tracking. *TIP*, 27:4585–4597, 2018.

[28] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High Performance Visual Tracking with Siamese Region Proposal Network. In *CVPR*, 2018.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dolí. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[30] Qiankun Liu, Qi Chu, Bin Liu, and Nenghai Yu. GSM: Graph Similarity Model for Multi-Object Tracking. In *IJCAI*, 2020.

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016.

[32] Andrii Maksai, Xinchao Wang, Francois Fleuret, and Pascal Fua. Non-markovian globally consistent multi-object tracking. In *ICCV*, 2017.

[33] Andrii Maksai, Xinchao Wang, and Pascal Fua. What players do with the ball: A physically constrained interaction modeling. In *CVPR*, 2016.

[34] Anton Milan, Laura Leal-Taixé, Taix´ Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A Benchmark for Multi-Object Tracking. Technical report, 2016.

[35] Anton Milan, Seyed Hamid Rezatofighi, Anthony Dick, Ian Reid, and Konrad Schindler. Online Multi-Target Tracking Using Recurrent Neural Networks. In *AAAI*, 2016.

[36] Songhwai Oh, Stuart Russell, and Shankar Sastry. Markov Chain Monte Carlo Data Association for Multiple-Target Tracking. *IEEE Transactions on Automatic Control*, 54(3):481–497, 2009.

[37] Jinlong Peng, Tao Wang, Weiyao Lin, Jian Wang, John See, Shilei Wen, and Erui Ding. TPM: Multiple object tracking with tracklet-plane matching. *Pattern Recognition*, 107:107480, 2020.

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Technical report, 2015.

[40] Weihong Ren, Xinchao Wang, Jiandong Tian, Yandong Tang, and Antoni B. Chan. Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets. *TIP*, 30:1439–1452, 2021.

[41] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint Probabilistic Data Association Revisited. In *ICCV*. IEEE, 2015.

[42] 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 workshop*, 2016.

[43] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking The Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies. In *ICCV*, 2017.

[44] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep Network Flow for Multi-Object Tracking. In *CVPR*, 2017.

[45] Han Shen, Lichao Huang, Chang Huang, and Wei Xu. Tracklet Association Tracker: An End-to-End Learning-based Association Approach for Multi-Object Tracking. Technical report, 2018.

[46] J. Shen, Z. Liang, J. Liu, H. Sun, L. Shao, and D. Tao. Multiobject tracking by submodular optimization. *IEEE Transactions on Cybernetics*, 49(6):1990–2001, 2019.

[47] Hao Sheng, Jiahui Chen, Yang Zhang, Wei Ke, Zhang Xiong, and Jingyi Yu. Iterative Multiple Hypothesis Tracking with Tracklet-level Association. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(8):1, 2018.

[48] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple People Tracking by Lifted Multicut and Person Re-identification. In *CVPR*, 2017.

[49] Lu Wang, Lisheng Xu, Min Young Kim, Luca Rigazico, and Ming Hsuan Yang. Online multiple object tracking via flow and convolutional features. In *ICIP*, 2017.

[50] Xinchao Wang, Vitaly Ablavsky, Horesh Ben Shitrit, and Pascal Fua. Take your eyes off the ball: Improving ball-tracking by focusing on team play. *CVIU*, 119:102–115, 2014.

[51] Xinchao Wang, Bin Fan, Shiyu Chang, Zhangyang Wang, Xianming Liu, Dacheng Tao, and Thomas S. Huang. Greedy batch-based minimum-cost flows for tracking multiple objects. *TIP*, 26:4765–4776, 2017.

[52] Xinchao Wang, Engin Turetken, Francois Fleuret, and Pascal Fua. Tracking interacting objects optimally using integer programming. In *ECCV*, 2014.

[53] Xinchao Wang, Engin Turetken, Francois Fleuret, and Pascal Fua. Tracking interacting objects using intertwined flows. *TPAMI*, 38:2312–2326, 2016.

[54] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple Online and Realtime Tracking with a Deep Association Metric. In *ICIP*, 2017.

[55] Jun Xiang, Guoshuai Zhang, Nong Sang, Rui Huang, and Jianhua Hou. Multiple Object Tracking by Learning Feature Representation and Distance Metric Jointly. In *BMVC*, 2018.

[56] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-Temporal Relation Networks for Multi-Object Tracking. In *ICCV*, 2019.

[57] 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*, 2020.

[58] 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*, 2016.

[59] Junbo Yin, Wenguan Wang, Qinghao Meng, Ruigang Yang, and Jianbing Shen. A Unified Object Motion and Affinity Model for Online Multi-Object Tracking. In *CVPR*, 2020.

[60] Guan Zhai, Huadong Meng, Zhiwen Zhong, and Xiqin Wang. A multiple hypothesis tracking method for extended target tracking. In *ICECE*, 2010.

[61] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction. In *CVPR*, 2019.

[62] Dawei Zhao, Hao Fu, Liang Xiao, Tao Wu, and Bin Dai. Multi-object Tracking with Correlation Filter for Autonomous Vehicle. *Sensors (Switzerland)*, 18(7):1–17, 2018.

[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 *ECCV*, 2018.