Looking Beyond Two Frames: End-to-End Multi-Object Tracking Using Spatial and Temporal Transformers

Tianyu Zhu[1,4], Markus Hiller[1], Mahsa Ehsanpour[3,4], Rongkai Ma[1,4], Tom Drummond[1,4], Hamid Rezatofighi[2]

[1]Department of Electrical and Computer Systems Engineering, Monash University
[2]Department of Data Science and AI, Monash University
[3]Australian Institute for Machine Learning, The University of Adelaide
[4]Australian Centre for Robotic Vision

tianyu.zhu@monash.edu

Abstract

Tracking a time-varying indefinite number of objects in a video sequence over time remains a challenge despite recent advances in the field. Ignoring long-term temporal information, most existing approaches are not able to properly handle multi-object tracking challenges such as occlusion. To address these shortcomings, we present MO3TR: a truly end-to-end Transformer-based online multi-object tracking (MOT) framework that learns to handle occlusions, track initiation and termination without the need for an explicit data association module or any heuristics/post-processing. MO3TR encodes object interactions into long-term temporal embeddings using a combination of spatial and temporal Transformers, and recursively uses the information jointly with the input data to estimate the states of all tracked objects over time. The spatial attention mechanism enables our framework to learn implicit representations between all the objects and the objects to the measurements, while the temporal attention mechanism focuses on specific parts of past information, allowing our approach to resolve occlusions over multiple frames. Our experiments demonstrate the potential of this new approach, reaching new state-of-the-art results on multiple MOT metrics for two popular multi-object tracking benchmarks. Our code will be made publicly available.

1. Introduction

Visually discriminating the identity of multiple objects in a scene and creating individual *tracks* of their movements over time, namely *multi-object tracking*, is one of the basic yet most crucial vision tasks, imperative to tackle many real-world problems in surveillance, robotics/autonomous driving, health and biology. While being a classical AI problem, it is still very challenging to design a reliable multi-object tracking (MOT) system capable of tracking an unknown and time-varying number of objects moving through unconstrained environments, directly from spurious and ambiguous measurements and in presence of many other complexities such as occlusion, detection failure and data (measurement-to-objects) association uncertainty.

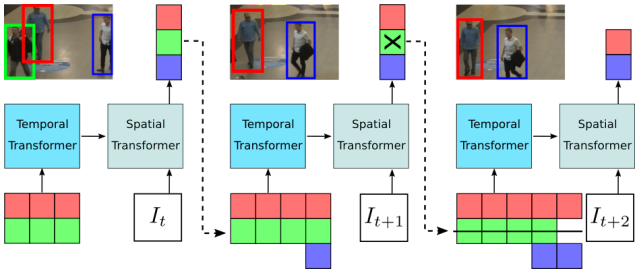

Figure 1. Looking beyond two frames with MO3TR: Temporal and spatial Transformers jointly pay attention to the current image I_t and the entire embedding history of the two tracked objects (*red* and *green*, left). Detection of a previously untracked object (*blue*) causes initiation of new track (left $\rightarrow$ middle), while an object exiting the scene (*green*) leads to track termination (middle $\rightarrow$ right). Embeddings encoding spatial and temporal interactions are accumulated over time to form individual object-track histories.

Early frameworks approached the MOT problem by splitting it into multiple sub-problems such as object detection, data association, track management and filtering/state prediction; each with their own set of challenges and solutions [1, 2, 6, 7, 19, 42, 51, 52]. Recently, deep learning has considerably contributed to improving the performance of multi-object tracking approaches, but surprisingly not through learning the entire problem end-to-end. Instead, the developed methods adopted the traditional problem split and mainly focused on enhancing some of the aforementioned components, such as creating better detectors [17, 38, 39, 40, 64] or developing more reliable matching objectives for associating detections to existing object

tracks [22, 29, 46, 58, 59]. While this tracking-by-detection paradigm has become the de facto standard approach for MOT, it has its own limitations. Recent approaches have shown advances by considering detection and tracking as a joint learning task rather than two separate sequential problems [4, 16, 54, 67]. However, these methods often formulate the MOT task as a two consecutive frames problem and ignore long-term temporal information, which is imperative for tackling key challenges such as track initiation, termination and occlusion handling.

In addition to their aforementioned limitations, all these methods can barely be considered to be end-to-end multi-object frameworks as their final outputs, *i.e*. tracks, are generated through a non-learning process. For example, track initiation and termination are commonly tackled by applying different heuristics, and the track assignments are decided upon by applying additional optimization methods, *e.g*. the Hungarian algorithm [26], max-flow min-cut [18], *etc*., and the generated tracks may be smoothed by a process such as interpolation or filtering [23].

With the recent rise in popularity of Transformers [56], this rather new deep learning tool has been adapted to solve computer vision problems like object detection [9] and, concurrent to our work, been deployed to two new MOT frameworks [33, 53]. Nonetheless, they still either rely on conventional heuristics, *e.g*. IoU matching [53], or formulate the problem as a two-frames task [33, 53], making them naive approaches to handle long-term occlusions.

In this paper, we will show that the MOT problem can be learnt end-to-end, without the use of heuristics or post-processing, addressing the key tasks like track initiation and termination, as well as occlusion handling. Our proposed method, nicknamed *MO3TR*, is a truly end-to-end Transformer-based online multi-object tracking method, which learns to recursively predict the state of the objects directly from an image sequence stream. Moreover, our approach encodes long-term temporal information to estimate the state of all the objects over time and does not contain an explicit data association module (Fig. 1).

Precisely speaking, MO3TR incorporates long-term temporal information by casting *temporal attention* over all past embeddings of each individual object, and uses this information to predict an embedding suited for the current time step. This access to longer-term temporal information beyond two frames is crucial in enabling the network to learn the difference between occlusion and termination, which is further facilitated through a specific data augmentation strategy. To factor in the influence of other objects and the visual input measurement, we refine the predicted object embedding by casting spatial attention over all identified objects in the current frame (*object-to-object attention*) as well as over the objects and the encoded input image (*object-to-image attention*).

The idea of this joint approach relates to the natural way humans perceive such scenarios: We expect certain objects to become occluded given their past trajectory and their surroundings, and predict when and where they will reappear.

To summarize, our main contributions are as follows:

1) We introduce an end-to-end tracking approach that learns to encode longer-term information beyond two frames through temporal and spatial Transformers, and recursively predicting all states of the tracked objects
2) We realize joint learning of object initialization, termination and occlusion handling without explicit data association and eliminate the need for heuristic post-processing
3) MO3TR reaches new state of the art results on two popular multi-object tracking benchmarks

2. Related work

Tracking-by-detection. Tracking-by-detection treats the multi-object tracking (MOT) task as a two-stage problem. Firstly, all objects in each frame are identified using an object detector [17, 39, 40, 64]. Detected objects are then associated over frames, resulting in tracks [6, 11]. The incorporation of appearance features and motion information has been proven to be of great importance for MOT. Appearance and ReID features have been extensively utilized to improve the robustness of multi-object tracking [25, 27, 29, 44, 63]. Further, incorporating motion has been achieved by utilizing a Kalman filter [23] to approximate the displacement of boxes between frames in a linear fashion and with the constant velocity assumption [1, 10] to associate detections [6, 59]. Recently, more complex and data-driven models have been proposed to model motion [15, 31, 66, 67] in a deterministic [37, 46] and probabilistic [15, 47, 57] manner. Graphs neural networks have been also used in the recent detection based MOT frameworks, conducive to extract a reliable global feature representation from visual and/or motion cues [8, 21, 50, 55].

Despite being highly interrelated, detection and tracking tasks are treated independently in this line of works. Further, the performance of tracking by detection methods highly relies on incorporating heuristics and post-processing steps to infer track initiation and termination, handle occlusions and assign tracks.

Joint detection and tracking. The recent trend in MOT has moved from associating detections over frames to regressing the previous track locations to new locations in the current frame. [4, 16, 67] perform temporal realignment by exploiting a regression head. Although detection and tracking are not disjoint components in these works, they still suffer from some shortcomings. These works formulate the problem as detection matching between two/few frames, thus solving the problem locally and ignoring long-term temporal information. We argue that MOT is a challenging task which requires long-term temporal encoding of object dy-

namics to handle object initiation, termination, occlusion and tracking. Furthermore, these approaches still rely on the conventional post processing steps and heuristics to generate the tracks.

Transformers for vision. Recently, Transformers [56] have been widely applied to many computer vision problems [3, 9, 35, 36], including MOT by two concurrent works [33, 53]. [53] performs multi-object tracking using a query-key mechanism which relies on heuristic post processing to generate final tracks. Trackformer [33] has been proposed as a transformer-based model which achieves joint detection and tracking by converting the existing DETR [9] object detector to an end-to-end trainable MOT pipeline. However, it still considers local information (two consecutive frames) to learn and infer tracks and ignores long-term temporal object dynamics, which are essential for effective learning of all MOT components.

This paper. To overcome all the existing limitations in the previous works, we propose an end-to-end MOT model which learns to jointly track multiple existing objects, handle their occlusion or terminate their tracks and initiate new tracks considering long-term temporal object information.

3. MO3TR

Learning an object representation that encodes both the object's own state over time and the interaction with its surroundings is vital to allow reasoning about three key challenges present in end-to-end multiple object tracking (MOT), namely *track initiation*, *termination* and *occlusion handling*. In this section, we demonstrate how such a representation can be acquired and continuously updated through our proposed framework: Multi-Object TRacking using spatial TRansformers and temporal TRansformers – short *MO3TR* (Fig. 2). We further introduce a training paradigm to learn resolving these three challenges in a joint and completely end-to-end trainable manner. We first present an overview of our framework and introduce the notation used throughout this paper, followed by a detailed introduction of the core components.

3.1. System overview and notation

The goal of tracking multiple objects in a video sequence of T frames is to retrieve an overall set of tracks $\mathbb{T}_T$ representing the individual trajectories for all uniquely identified objects present in at least one frame. Given the first frame I_0 at time t_0, our model tentatively initializes a set of tracks $\mathbb{T}_0$ based on all objects identified for this frame. From the next time step onward, the model aims to compute a set of embeddings $\mathcal{Z}_t = \{\boldsymbol{z}_t^1, \boldsymbol{z}_t^2, \dots, \boldsymbol{z}_t^M\}$ representing all M objects present in the scene at time t (Fig. 2). Taking in the track history $\mathbb{T}_{t-1}$ from the previous time step, we predict a set of embeddings $\hat{\mathcal{Z}}_t$ for the current time step based on the past representations of all objects using temporal attention (Section 3.2). Together with a learnt set of representation queries $\mathcal{Z}_Q$ proposing the initiation of new object tracks, these predicted object representations are processed by our first spatial attention module to reason about the interaction occurring between different objects (Section 3.3). This refined set of intermediate object representations $\mathcal{Z}_t'$ is then passed to the second spatial attention module which takes the interaction between the objects and the scene into account by casting attention over the object embeddings and the visual information of the current frame I_t transformed into its feature map $\boldsymbol{x}_t$ (Section 3.3). This two-step incorporation of spatial information into the embeddings is iteratively performed multiple times over several layers, returning the final set of refined object representations $\mathcal{Z}_t$.

The incorporation of temporal and spatial information into a representative embedding of any object m at time t

$$\boldsymbol{z}_t^m = f(\mathbb{T}_{t-1}, \mathcal{Z}_Q, \boldsymbol{x}_t) \tag{1}$$

can be summarized as a learnt function $f(\cdot)$ of the track history $\mathbb{T}_{t-1}$, the learnt set of initiation queries $\mathcal{Z}_Q$ and the encoded image feature map $\boldsymbol{x}_t$. This function representation demonstrates our main objective to enable the framework to learn the best possible way to relate the visual input to the objects' internal states, without enforcing overly-restrictive constraints or explicit data association.

The use of the resulting embeddings $\mathcal{Z}_t$ in our framework is twofold. Tracking results in the form of object-specific class scores $\boldsymbol{c}_t^m$ and corresponding bounding boxes $\boldsymbol{b}_t^m$ for the current frame are obtained through simple classification and bounding box regression networks (Fig. 2). Further, the subset of embeddings yielding a high probability of representing an object present in the current frame ($p_{\boldsymbol{z}_t^m}(c_{\text{obj}}) > 0.5$) is added to the track history to form the basis for the prediction performed in the next time step. Throughout the entire video sequence, new tracks $\mathcal{T}_{s_m}^m$ representing objects that enter the scene are initialized, while previous tracks may be terminated for objects no longer present. This leads to an overall set of tracks $\mathbb{T}_T = \{\mathcal{T}_{s_1:e_1}^1, \dots, \mathcal{T}_{s_N:e_N}^N\}$ for all N uniquely identified objects present in at least one frame of the video sequence of length T, with their life span indicated by the subscript as initiation (*start*) and termination (*end*) time, respectively.

3.2. Learning long-term temporal embeddings

Discerning whether an object is not visible in a given frame due to occlusion or because it is no longer present in the scene is challenging. Considering that visual features extracted during partial or full occlusion are not describing the actual object they aim to represent increases this even further. Humans naturally reach decisions in such scenarios by considering all available information jointly. Analyzing the motion behavior of objects up to that point, we ignore frames with non-helpful information, and predict how and where the object is expected to re-appear in the current frame. Intuitively, MO3TR follows a similar approach.

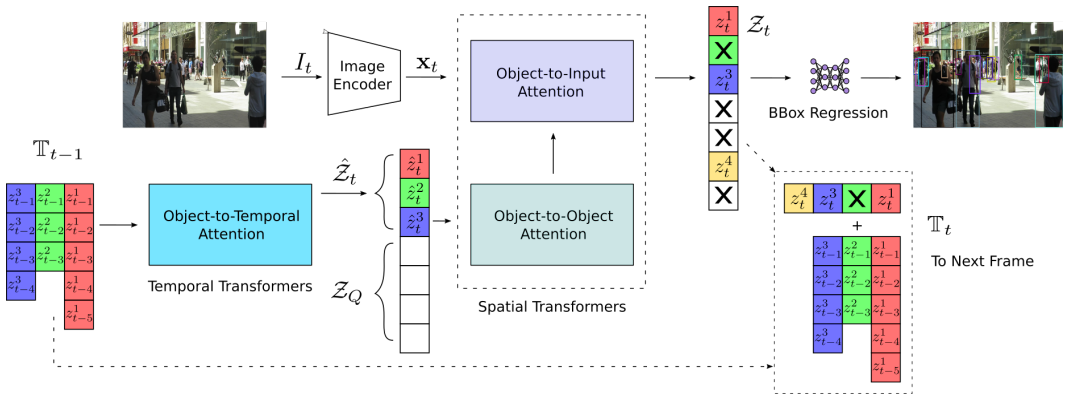

Figure 2. Overview of our MO3TR framework. Starting from the left, the temporal Transformer uses the entire embedding-based track history $\mathbb{T}_{t-1}$ to predict representative object encodings $\hat{\mathcal{Z}}_t$ for the current, yet unobserved, time step t. The spatial Transformer then jointly considers the predictions together with a set of learnt initiation embeddings $\mathcal{Z}_Q$ and the input image I_t to reason about all objects in a joint manner, determining the initiation of new and termination of existing tracks. Embeddings of identified objects in $\mathcal{Z}_t$ are used to regress corresponding bounding boxes describing the tracked objects, and are appended to form the track history $\mathbb{T}_t$ for the next frame.

Our framework learns the temporal behavior of objects jointly with the rest of the model through a Transformer-based component [56] that we nickname *temporal Transformer*. For any tracked object m at time t, the temporal Transformer casts attention over all embeddings contained in the object's track history $\mathcal{T}^m_{t-1} = \{\boldsymbol{z}^m_{s_m}, ..., \boldsymbol{z}^m_{t-1}\}$, and predicts a thereon-based *expected* object representation $\hat{\boldsymbol{z}}^m_t$ for the current frame. We supplement each object's track history $\mathcal{T}^m_{t-1}$ by adding positional encodings [56] to the embeddings in the track to represent their relative time in the sequence. We denote the time-encoded track history by $\mathcal{T}^{m,\text{pe}}_{t-1}$ and individual positional time-encodings for time t as $pe_t \in \mathbb{R}$. Passing the *request* for an embedding estimate of the current time step t in form of the positional time-encoding pe_t as a query to the Transformer[1] and providing $\mathcal{T}^{m,\text{pe}}_{t-1}$ as basis for keys and values, we retrieve the predicted object embedding

$$\hat{\boldsymbol{z}}^m_t = \Psi\left(\frac{1}{\sqrt{d_{\boldsymbol{z}}}}\, q^{\text{tp}}(pe_t)\;\; k^{\text{tp}}\big(\mathcal{T}^{m,\text{pe}}_{t-1}\big)^{\text{T}}\right)\;\; v^{\text{tp}}\big(\mathcal{T}^{m,\text{pe}}_{t-1}\big)\,, \quad (2)$$

where Ψ represents the $\mathrm{softmax}$ operator, $q^{\text{tp}}(\cdot)$, $k^{\text{tp}}(\cdot)$ and $v^{\text{tp}}(\cdot)$ are learnt query, key and value functions of the temporal Transformer, respectively, and $d_{\boldsymbol{z}} \in \mathbb{R}$ denotes the dimension of the object embeddings.

In other words, the predicted representation $\hat{\boldsymbol{z}}^m_t$ of object m is computed through a dynamically weighted combination of all its previous embeddings. This allows the temporal Transformer to: (i) incorporate helpful and ignore irrelevant or faulty information from previous time steps, and (ii) predict upcoming occlusions and create appropriate embeddings that focus more on conveying important positional rather than visual information. While these tasks resemble those usually performed via heuristics and manual parameter tuning during track management, MO3TR learns these dependencies end-to-end without the need of heuristics.

In practice, the prediction of object representations introduced for the example of one tracked object in (2) is performed in a batched-parallel manner for the entire set of existing tracks $\mathbb{T}_{t-1}$ over multiple layers, resulting in the output set $\hat{\mathcal{Z}}_t$ of the temporal Transformers that is passed as input to the spatial Transformers (Fig. 2). Note that the size of the set is dynamic and depends on the number of tracked objects. Details on how the temporal Transformer is trained are provided in Section 3.4.

3.3. Learning spatial interactions

Multiple pedestrians that are present in the same environment not only significantly influence each others movements, but also their respective visual appearance through occluding each other when perceived from a fixed viewpoint. In this section, we introduce how MO3TR learns to incorporate these dependencies into the object representations. Starting from how detection and track initiation are performed within the concept of Transformers, we then detail the refinement of object embeddings by including the interaction between objects, followed by the interaction between objects and the input image.

Initiation of new tracks. For a new and previously untracked object m spawning at any time t, a corresponding track history $\mathcal{T}^m_{t-1}$ does not yet exist and hence, no predicted embedding is passed from the temporal to the spatial Transformer (Fig. 2). To allow initiation of new tracks for such detected objects, we build upon [9] and learn a fixed set of initiation queries $\mathcal{Z}_Q$. Intuitively, these queries learn to propose embeddings that lead the spatial Transformer to check for objects with certain properties and at certain locations in the visual input data. Importantly, these queries are con-

[1] Note that this method allows to predict embeddings for any future time step, and could thus be easily extended to further applications like trajectory forecasting, or similar.

sidered jointly with the ones propagated from the temporal Transformer to avoid duplicate tracks.

Interaction between tracked objects. We use self-attention [56] to capture the influence tracked objects have onto each other's motion behavior and appearance. This interaction aspect is incorporated into the object embeddings by computing an updated version of the representation set

$$\mathcal{Z}_t^{'} = \Psi\left(\frac{1}{\sqrt{d_{\boldsymbol{z}}}}\, q^{\text{sf}}(\bar{\mathcal{Z}}_t)\; k^{\text{sf}}(\bar{\mathcal{Z}}_t)^{\text{T}}\right) v^{\text{sf}}(\bar{\mathcal{Z}}_t), \tag{3}$$

where $q^{\text{sf}}(\cdot)$, $k^{\text{sf}}(\cdot)$ and $v^{\text{sf}}(\cdot)$ are all learnt functions of the concatenated object embedding set $\bar{\mathcal{Z}}_t = \{\hat{\mathcal{Z}}_t, \mathcal{Z}_Q\}$, $d_{\boldsymbol{z}}$ is the dimension of the embeddings and Ψ the $\mathrm{softmax}$ operator. Relating this approach to the classical transformer formulation, the functions conceptually represent the queries, keys and values introduced in [56].

Interaction between objects and the input image. The relationship between the set of objects and the image is modeled through encoder-decoder attention (*aka* cross-attention) to relate all object representations to the encoded visual information of the current image (*i.e.* measurement). Evaluating this interaction results in the computation of a second update to the set of object representations

$$\mathcal{Z}_t^{''} = \Psi\left(\frac{1}{\sqrt{d_{\boldsymbol{z}}}}\, q^{\text{cr}}(\mathcal{Z}_t^{\prime})\; k^{\text{cr}}(\boldsymbol{x}_t)^{\text{T}}\right) v^{\text{cr}}(\boldsymbol{x}_t), \tag{4}$$

where $q^{\text{cr}}(\cdot)$ is a learnt function of the pre-refined object embeddings $\mathcal{Z}_t^{\prime}$, and $k^{\text{cr}}(\cdot)$ and $v^{\text{cr}}(\cdot)$ are learnt functions of the image embedding $\boldsymbol{x}_t$ produced by a CNN backbone and a Transformer encoder. Ψ represents the $\mathrm{softmax}$ operator.

Combining interactions for refined embeddings. In practice, the two previously described update steps are performed consecutively with (4) taking as input the result of (3), and are iteratively repeated over several layers of the Transformer architecture. This sequential incorporation of updates into the representation is inspired by DETR [9], where self-attention and cross-attention modules are similarly deployed in a sequential manner. Using both introduced concepts of object-to-object and object-to-measurement attention allow the model to globally reason about all tracked objects via their pair-wise relationships, while using the current image as context information to retrieve the final set of updated object representations $\mathcal{Z}_t$.

Updating the track history. After each frame is processed by the entire framework, the final set of embeddings $\mathcal{Z}_t$ of objects identified to be present in the frame is added to the track history $\mathbb{T}_{t-1}$, creating the basis for the next prediction of embeddings by the temporal Transformer (Fig. 2). We consistently append new embeddings from the right-hand side, followed by right-aligning the entire set of embeddings. Due to the different lengths of tracks for different objects, this procedure aligns embeddings representing identical time steps, a method that we found to help stabilize training and improve the inference of the temporal Transformer (Table 4).

3.4. Training MO3TR

The training procedure of MO3TR (Fig. 2) is composed of two key tasks: (i) creating a set of suitable tracklets that can be used as input $\mathbb{T}_{t-1}$ to the temporal Transformer, and (ii) assigning the predicted set of M output embeddings $\mathcal{Z}_t = \{\boldsymbol{z}_t^m\}_{m=1}^M$ to corresponding ground truth labels of the training set, and applying a corresponding loss to facilitate training. With the number output embeddings being by design larger than the number of objects in the scene, matching occurs wither either with trackable objects or the background class.

Constructing the input tracklet set. The input to the model at any given time t is defined as the track history $\mathbb{T}_{t-1}$ and the current image I_t. To construct a corresponding $\mathbb{T}_{t-1}$ for any I_t sampled from the dataset during training, we first extract the ordered set of K directly preceding images $\{I_k\}_{k=t-K}^{t-1}$ from the training sequence. Passing these images without track history to MO3TR causes the framework to perform track initiation for all identified objects in each frame by using the trainable embeddings $\mathcal{Z}_Q$, returning an ordered set of output embedding sets $\{\mathcal{Z}_k\}_{k=t-K}^{t-1}$. Each output embedding set $\mathcal{Z}_k$ contains a variable number of M_k embeddings representing objects in the respective frame k. We use multilayer perceptrons (MLPs) to extract corresponding bounding boxes $\hat{\boldsymbol{b}}_k^m$ and class scores $\hat{\boldsymbol{c}}_k^m$ from each of these object embeddings $\boldsymbol{z}_k^m \in \mathcal{Z}_k$, resulting in a set of M_k object-specific pairs denoted as $\{\hat{y}_k^m\}_{m=1}^{M_k} = \{(\hat{\boldsymbol{b}}_k^m, \hat{\boldsymbol{c}}_k^m)\}_{m=1}^{M_k}$ for each frame k. The pairs are then matched with the ground truth $\{y_k^i\}_{i=1}^{G_k}$ of the respective frame through computing a bipartite matching [9] between these sets. The permutation $\hat{\sigma}_k$ of the M_k predicted elements with lowest pair-wise matching cost $\mathcal{C}_{\text{matching}}$ is determined by solving the assignment problem

$$\hat{\sigma}_k = \arg\min_{\sigma \in \mathcal{S}} \sum_i^{M_k} \mathcal{C}_{\text{matching}}\left(y_k^i, \hat{y}_k^{\sigma(i)}\right), \tag{5}$$

through the Hungarian algorithm [26], with the matching cost taking both the probability of correct class prediction $\hat{p}_k^{\sigma(i)}(\boldsymbol{c}_k^i)$ and bounding box similarity into account

$$\mathcal{C}_{\text{matching}} = -\hat{p}_k^{\sigma(i)}\left(\boldsymbol{c}_k^i\right) + \mathcal{C}_{\text{bbox}}\left(\boldsymbol{b}_k^i, \hat{\boldsymbol{b}}_k^{\sigma(i)}\right). \tag{6}$$

We follow [9] and use a linear combination of L1 distance and the scale-invariant generalized intersection over union [41] cost $\mathcal{C}_{\text{giou}}$ to mitigate any possible scale issues arising from different box sizes. The resulting bounding box cost with weights $\alpha_{\text{L1}}, \alpha_{\text{giou}} \in \mathbb{R}^+$ is then defined as

$$\mathcal{C}_{\text{bbox}} = \alpha_{\text{L1}} \left\| \boldsymbol{b}_k^i - \hat{\boldsymbol{b}}_k^{\sigma(i)} \right\|_1 + \alpha_{\text{giou}} \mathcal{C}_{\text{giou}}\left(\boldsymbol{b}_k^i, \hat{\boldsymbol{b}}_k^{\sigma(i)}\right). \tag{7}$$

The identified minimum cost matching between the output and ground truth sets is used to assign all embeddings classified as objects their respective identities annotated in the ground truth labels. The objects of all K frames are accumulated, grouped regarding their assigned identities and sorted in time-ascending order to form the overall set of previous object tracks $\mathbb{T}_{t-1}$ serving as input to our model.

Losses. Given the created input set of tracks $\mathbb{T}_{t-1}$ and the image I_t, MO3TR predicts an output set of object embeddings $\mathcal{Z}_t = \{\boldsymbol{z}_t^1, \boldsymbol{z}_t^2, \dots, \boldsymbol{z}_t^M\}$ at time t. Similar to before, we extract bounding boxes and class scores for each embedding in the set. However, embeddings that possess a track history already have unique identities associated to them and are thus directly matched with the respective ground truth elements. Only newly initiated embeddings without track history are then matched with remaining unassigned ground truth labels as previously described. Elements that could not be matched are assigned the *background* class. Finally, we re-use (6) and (7) for $k = t$ and apply them as our loss to the matched elements of the output set.

Data augmentation. Most datasets are highly imbalanced regarding the occurrence of occlusion, initiation and termination scenarios. To facilitate learning of correct tracking behaviour, we propose to mitigate the imbalance problem by modelling similar effects through augmentation:

1. We randomly drop a certain number of embeddings in the track history to simulate cases where the object could not be identified for some frames, aiming to increase robustness. If the most recent embedding is dropped, the model can learn to re-identify objects.
2. Random false positive examples are inserted into the history to simulate false detection and faulty appearance information due to occlusion. This aims for the model to learn ignoring unsuited representations through its attention mechanism.
3. We randomly select the sequence length used to create the track history during training to increase the model's capability to deal with varying track lengths.

The high importance of these augmentations are proved in Section 4.3 and Table 4.

4. Experiments

In this section, we demonstrate the performance of MO3TR by comparing against other multi-object tracking methods on popular MOT benchmarks[2] and evaluate different aspects of our contribution in detailed ablation studies. We further provide implementation and training details.

Datasets. We use the MOT16 and MOT17 [34] datasets from the MOTchallenge benchmarks to evaluate and compare MO3TR with other state of the art models. Both datasets contain seven training and test sequences each, capturing crowded indoor or outdoor areas via moving and static cameras from various viewpoints. Pedestrians are often heavily occluded by other pedestrians or background objects, making identity-preserving tracking challenging. Three sets of public detections are provided with MOT17 (DPM [17], FRCNN [40] and SDP [64]), and one with MOT16 (DPM). For ablation studies, we combine sequences of the new MOT20 benchmark [13] and 2DMOT15 [30] to form a diverse validation set covering both indoor and outdoor scenes at various pedestrian density levels.

[2] https://motchallenge.net/

Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
EAMTT [48]	38.8	42.4	7.9	49.1	8,114	102,452	965
DMAN [68]	46.1	54.8	17.4	42.7	7,909	89,874	**532**
AMIR [46]	47.2	46.3	14.0	41.6	2,681	92,856	774
MOTDT17 [32]	47.6	50.9	15.2	38.3	9,253	85,431	792
STRN [61]	48.5	53.9	17.0	34.9	9,038	84,178	747
UMA [65]	50.5	52.8	17.8	33.7	7,587	81,924	685
Tracktor++ [4]	54.4	52.5	19.0	36.9	3,280	79,149	682
Tracktor++v2 [4]	56.2	54.9	20.7	35.8	**2,394**	76,844	617
DeepMOT-T [62]	54.8	53.4	19.1	37.0	2,955	78,765	645
MO3TR (Res50)	**64.2**	**60.6**	**31.6**	**18.3**	7,620	**56,761**	929

Table 1. Results on the MOT16 benchmark [34] test set using public detections. **Bold** and underlined numbers indicate best and second best result, respectively. More detailed results of our approach are provided in the supplementary material.

Evaluation metrics. To evaluate our model and other MOT methods, we use standard metrics recognized by the tracking community [5, 43]. The two main metrics are the MOT Accuracy (MOTA) and Identity F1 Score (IDF1). MOTA focuses more on object coverage while the consistency of assigned identities is measured by IDF1. We further report False Positives (FP), False Negatives (FN), Mostly Tracked (MT) and Mostly Lost (ML). Further details of these metrics are provided in the supplementary material.

4.1. Implementation details of MO3TR

We employ a multi-stage training concept to train MO3TR end-to-end. Firstly, our ImageNet [45] pretrained ResNet50 [20] backbone is, together with the encoder and spatial Transformers, trained on a combination of the CrowdHuman [49], ETH [14] and CUHK-SYSU [60] datasets for 300 epochs on a pedestrian detection task. This training procedure is similar to DETR [9]. Afterwards, we engage our temporal transformer and train the entire model end-to-end using the MOT17 dataset for another 300 epochs. The initial learning rate for both training tasks is 1e-4, and is dropped by a factor of 10 every 100 epochs. Relative weights of our loss are the same as in DETR [9], the number of initiation queries is 100. The input sequence length representing object track histories varies randomly from 1 to 30 frames. To enhance the learning of temporal encoding, we predict 10 future frames instead of one and compute the total loss. We train our model using 4 GTX

Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
GMPHD [28]	39.6	36.6	8.8	43.3	50,903	284,228	5,811
EAMTT [48]	42.6	41.8	12.7	42.7	30,711	288,474	4,488
SORT17 [6]	43.1	39.8	12.5	42.3	28,398	287,582	4,852
DMAN [68]	48.2	55.7	19.3	38.3	26,218	263,608	2,194
MOTDT17 [32]	50.9	52.7	17.5	35.7	24,069	250,768	2,474
STRN [61]	50.9	56.5	20.1	37.0	27,532	246,924	2,593
jCC [24]	51.2	54.5	20.9	37.0	25,937	247,822	**1,802**
DeepMOT-T [62]	53.7	53.8	19.4	36.6	11,731	247,447	1,947
FAMNet [12]	52.0	48.7	19.1	33.4	14,138	253,616	3,072
UMA [65]	53.1	54.4	21.5	31.8	22,893	239,534	2,251
Tracktor++ [4]	53.5	52.3	19.5	36.6	12,201	248,047	2,072
Tracktor++v2 [4]	56.5	55.1	21.1	35.3	**8,866**	248,047	3,763
CenterTrack[67]	61.5	59.6	26.4	31.9	14,076	200,672	2,583
Trackformer [33]	61.8	59.8	**35.4**	21.1	35,226	**177,270**	2,982
MO3TR (Res50)	**63.2**	**60.2**	31.9	**19.2**	21,966	182,860	2,841

Table 2. Results on the MOT17 benchmark [34] test set using public detections. **Bold** and underlined numbers indicate best and second best result, respectively. More detailed results of our approach are provided in the supplementary material.

1080ti GPUs with 11GB memory each. It is to be noted that these computational requirements are significantly lower than for other recently published approaches in this field. We expect the performance of our model to further increase through bigger backbones and longer sequence length as well as an increased number of objects per frame.

Public detection. We evaluate the tracking performance using the public detections provided by the MOTChallenge. Not being able to directly produce tracks from these detections due to being an embedding-based method, we follow [33, 67] in filtering our initiations by the public detections using bounding box center distances, and only allow initiation of matched and thus publicly detected tracks.

4.2. Comparison with the state of the art

We evaluate MO3TR on the challenging MOT16 [34] and MOT17 benchmark test datasets using the provided public detections and report our results in Tables 1 and 2, respectively. Despite not using any heuristic track management to filter or post-process, we outperform most competing methods and achieve new state of the art results on both datasets regarding MOTA, IDF1 and ML metrics, and set a new benchmark for MT and FN on MOT16.

As clearly shown by its state of the art IDF1 scores on both datasets, MO3TR is capable of identifying objects and maintaining their identities over long parts of the track, in many cases for more than 80% of the objects' lifespans as evidenced by the very high MT results. Access to the track history through the temporal Transformers and jointly reasoning over existing tracks, initiation and the input data through the spatial Transformers helps MO3TR to learn discerning occlusion from termination. The framework is thus capable to avoid false termination, as clearly evidenced by the very low FN and record low ML numbers achieved on both MOT datasets. These values further indicate that MO3TR learns to fill in gaps due missed detections or occlusions, which has additional great influence on reducing FN and ML while increasing IDF1 and MT. Using its joint reasoning over the available information helps MO3TR to reduce failed track initiations (FN) considerably while keeping incorrect track initiations (FPs) at a reasonable low levels. The combination of superior IDF1, very low FN and reasonable FP allows MO3TR to reach new state of the art MOTA results on both MOT16 (Table 1) and MOT17 (Table 2) datasets.

len($\mathcal{T}^m_{t-1}$)	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓
1	55.4	48.4	114	19	4,700	12,898
10	56.8	49.0	115	**18**	4,245	12,805
20	57.8	50.1	115	19	3,826	12,787
30	**58.9**	**50.6**	114	20	**3,471**	**12,692**

Table 3. Effect of varying lengths of track history $\mathcal{T}^m_{t-1}$ considered in the temporal Transformers during evaluation.

Training Strategies	MOTA↑	IDF1↑	FP↓	FN↓
Naive (Two Frames)	12.2	22.1	7,905	26,848
FN (Two Frames)	14.6	42.0	22,609	**11,671**
FN+RA (Two Frames)	28.4	42.5	16,749	11,940
FN+RA+FP (Two Frames)	55.4	48.4	3,927	17,912
FN	21.9	42.5	19,353	11,693
FN+RA	39.2	48.1	12,265	12,002
FN+RA+FP	**58.9**	**50.6**	**3,471**	12,692

Table 4. Effect of different training (two frames vs. 30) and augmentation strategies: False Negatives (FN), False Positives (FP), Right-Aligned insertion (RA).

4.3. Ablation studies

In this section, we evaluate different components of MO3TR on our validation set using private detections and show the individual contributions of the key components and strategies to facilitate learning.

Effect of track history length. The length of the track history describes the maximum number of embeddings from all the previous time steps of a certain identified object that our temporal Transformer has access to. To avoid overfitting to any particular history length that might be dominant in the dataset but not actually represent the most useful source of information, we specifically train our model with input track histories of varying and randomly chosen lengths. It is important to note that if the maximum track history length is set to one, the method practically degenerates to a two-frame based joint detection and tracking method such as Trackformer [33]. Our results reported in Table 3 however show that incorporating longer-term information is crucial to improve end-to-end tracking. Both MOTA and IDF1 can be consistently improved while FP can be reduced when longer term history, *i.e.*, information from previous frames, is taken into account. This trend is also

Figure 3. Qualitative results of two challenging occlusion scenarios in the validation set. Objects of focus are highlighted with slightly thicker bounding boxes. Unlike Tracktor++v2 [4], our proposed MO3TR is capable of retaining the identity and keeps track even if the object is severely occluded.

clearly visible throughout evaluation of our training strategies presented in Table 4, further discussed in the following.

Training strategies. MOT datasets are highly imbalanced when it comes to the occurrence of initialization and termination examples compared to normal propagation, making it nearly impossible for models to naturally learn initiation of new or termination of no longer existing tracks when trained in a naive way. As presented in Table 4, naive training without any augmentation shows almost double the number of false negatives (FN) compared to augmented approaches, basically failing to initiate tracks properly. Augmenting with FN as discussed in 3.4 shows significant improvements for both two-frame and longer-term methods. Additionally right-aligning the track history helps generally to stabilize training and greatly reduces false positives. At last, augmenting with false positives is most challenging to implement but crucial. As the results demonstrate, it significantly reduces false positives by helping the network to properly learn the terminating of tracks.

Analysing temporal attention. To provide some insight into the complex and highly non-linear working principle of our temporal Transformers, we visualize the attention weights over the temporal track history for different track history lengths averaged for 100 randomly picked objects in our validation set (Fig. 4). Results for the first layer clearly depict most attention being payed to multiple of its more recent frames, decreasing with increasing frame distance. The second and third layers are harder to interpret due to the increasing non-linearity, and the model starts to increasingly cast attention over more distant frames. It is important

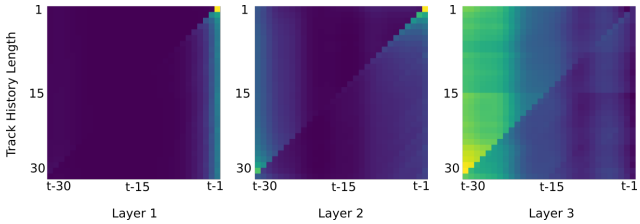

Figure 4. Temporal attention maps averaged over 100 randomly selected objects from the MOT20 dataset [13]. The vertical axis represents the maximum track history length, the horizontal axis the different embedding positions in the history. The displayed attention related the current query at time t to all the previous embeddings. Every row sums up to 1.

to notice that even if an embedding is not available at time $t-k$, the model can still choose to pay attention to that slot and use the non-existence for reasoning.

5. Conclusion

We presented MO3TR, a truly end-to-end multi-object tracking framework that uses temporal Transformers to encode the history of objects while employing spatial Transformers to encode the interaction between objects and the input data, allowing it to handle occlusions, track termination and initiation. Demonstrating the advantages of long term temporal learning, we set new state of the art results regarding multiple metrics on the popular MOT16 and MOT17 benchmarks.

A. Experiments

In this section, we provide details on the evaluation metrics used throughout the main paper, as well as detailed results for all sequences on the MOT16 and MOT17 benchmarks [34].

A.1. Evaluation metrics

To evaluate MO3TR and compare its performance to other state-of-the-art tracking approaches, we use the standard set of metrics proposed in [5, 43]. Analyzing the detection performance, we provide detailed insights regarding the total number of *false positives* (FP) and *false negatives* (FN, *i.e.* missed targets). The *mostly tracked targets* (MT) measure describes the ratio of ground-truth trajectories that are covered for at least 80% of the track's life span, while *mostly lost targets* (ML) represents the ones covered for at most 20%. The number of *identity switches* is denoted by IDs. The two most commonly used metrics to summarize the tracking performance are the *multiple object tracking accuracy* (MOTA), and the identity F1 score (IDF1). MOTA combines the measures for the three error sources of false positives, false negatives and identity switches into one compact measure, and a higher MOTA score implies better performance of the respective tracking approach. The IDF1 represents the ratio of correctly identified detections over the average number of ground-truth and overall computed detections.

All reported results are computed by the official evaluation code of the MOTChallenge benchmark[3].

A.2. Evaluation results

The public results for the MOT16 [34] benchmark presented in the experiment section of the main paper show the overall result of MO3TR on the benchmark's test dataset using the provided public detections (DPM [17]). Detailed results showing the results for all individual sequences are presented in Table A1. Similarly the individual results for all sequences of the MOT17 benchmark [34] comprising three different sets of provided public detections (DPM [17], FRCNN [40] and SDP [64]) are detailed in Table A2. Further information regarding the metrics used is provided in Section A.1.

B. Data association as auxiliary task

In the introduction of the main paper, we introduce the idea that our proposed MO3TR performs tracking *without any explicit data association module*. To elaborate what we mean by that and how multi-object tracking (MOT) without an explicitly formulated data association task is feasible, we would like to re-consider the actual definition of the MOT problem: Finding a mapping from any given input data, *e.g.* an image sequence stream, to the output data, *i.e.* a set of object states over time. In any learning scheme, given a suitable learning model, this mapping function can theoretically be learned without the requirement for solving any additional auxiliary task, as long as the provided inputs and outputs are clearly defined. The firmly established task of data association, *e.g.* a minimum cost assignment (*e.g.* using Hungarian Algorithm) between detections and objects, is nothing more than such an auxiliary task originally created to solve tracking based on tracking-by-detection paradigms. An end-to-end learning model, however, can learn to infer implicit correspondences and thus renders the explicit formulation of this task obsolete.

Precisely speaking, our end-to-end tracking model learns to relate the visual input information to the internal states of the objects via a self-supervised attention scheme. We realize this through using a combination of Transformers [56] to distill the available spatial and temporal information into representative object embeddings (*i.e.* the object states), making the explicit formulation of any auxiliary data association strategy unnecessary.

[3] https://motchallenge.net

Sequence	Public detector	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
MOT16-01	DPM [17]	62.83	57.29	7	1	99	2,251	27
MOT16-03	DPM	73.75	67.66	77	15	2,850	24,435	166
MOT16-06	DPM	59.52	58.74	81	41	1,083	3,442	146
MOT16-07	DPM	57.90	51.02	14	8	468	6,286	118
MOT16-08	DPM	44.69	39.91	13	13	489	8,610	158
MOT16-12	DPM	48.92	59.86	26	17	1,382	2,806	49
MOT16-14	DPM	43.49	44.88	22	44	1,249	8,931	265
MOT17	**All**	**64.18**	**60.59**	**240**	**139**	**7,620**	**56,761**	**929**

Table A1. Detailed MO3TR results on each individual sequence of the MOT16 benchmark [34] test set using public detections. Following other works, we use the public detection filtering method using center distances as proposed by [4].

Sequence	Public detector	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
MOT17-01	DPM [17]	62.31	57.00	6	2	98	2,306	27
MOT17-03	DPM	73.82	67.73	78	15	2,773	24,469	167
MOT17-06	DPM	60.51	58.81	81	39	950	3,555	148
MOT17-07	DPM	56.70	50.47	13	12	402	6,793	120
MOT17-08	DPM	38.00	35.42	13	26	206	12,723	168
MOT17-12	DPM	49.34	59.72	28	20	1,268	3,073	50
MOT17-14	DPM	43.49	44.88	22	44	1,249	8,931	265
MOT17-01	FRCNN [40]	60.37	53.92	7	4	109	2,419	28
MOT17-03	FRCNN	73.88	68.13	75	15	3,036	24,148	161
MOT17-06	FRCNN	61.98	61.07	95	25	1,170	3,148	162
MOT17-07	FRCNN	56.70	50.63	12	12	402	6,794	118
MOT17-08	FRCNN	36.48	35.17	12	31	149	13,121	149
MOT17-12	FRCNN	50.96	61.36	26	26	970	3,239	41
MOT17-14	FRCNN	43.67	44.08	24	42	1,349	8,790	272
MOT17-01	SDP [64]	66.76	57.60	7	3	94	2,022	28
MOT17-03	SDP	74.07	68.15	80	15	3,373	23,606	163
MOT17-06	SDP	61.91	61.21	93	24	1,163	3,176	150
MOT17-07	SDP	57.38	50.24	13	12	407	6,672	120
MOT17-08	SDP	38.62	36.32	13	27	235	12,553	177
MOT17-12	SDP	50.43	59.66	29	20	1,200	3,043	53
MOT17-14	SDP	46.35	45.94	24	38	1,363	8,279	274
MOT17	**All**	**63.19**	**60.15**	**751**	**452**	**21,966**	**182,860**	**2,841**

Table A2. Detailed MO3TR results on each individual sequence of the MOT17 benchmark [34] test set using public detections. Following other works, we use the public detection filtering method using center distances as proposed by [4].

References

[1] Anton Andriyenko and Konrad Schindler. Multi-target tracking by continuous energy minimization. In *CVPR*, 2011. 1, 2

[2] Anton Andriyenko, Konrad Schindler, and Stefan Roth. Discrete-continuous optimization for multi-target tracking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1926–1933. IEEE, 2012. 1

[3] Irwan Bello, Barret Zoph, Quoc Le, Ashish Vaswani, and Jonathon Shlens. Attention augmented convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3285–3294, 2019. 3

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 2, 6, 7, 8, 10

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6, 9

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468, 2016. 1, 2, 7

[7] Samuel S. Blackman and Robert Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House radar library. Artech House, 1999. 1

[8] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020. 2

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020. 2, 3, 4, 5, 6

[10] Wongun Choi and Silvio Savarese. Multiple target tracking in world coordinate with single, minimally calibrated camera. In *European Conference on Computer Vision*, 2010. 2

[11] Peng Chu, Heng Fan, Chiu C Tan, and Haibin Ling. Online multi-object tracking with instance-aware tracker and dynamic model refreshment. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 161–170, 2019. 2

[12] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6172–6181, 2019. 7

[13] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 6, 8

[14] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 6

[15] Kuan Fang, Yu Xiang, Xiaocheng Li, and Silvio Savarese. Recurrent autoregressive networks for online multi-object tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2

[16] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[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, 2009. 1, 2, 6, 9, 10

[18] Lester R. Ford and Delbert R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956. 2

[19] T. Fortmann, Y. Bar-Shalom, and M. Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE Journal of Oceanic Engineering*, 8(3):173–184, 1983. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[21] Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, and Paul Swoboda. Lifted disjoint paths with application in multiple object tracking. In *International Conference on Machine Learning*, pages 4364–4375, 2020. 2

[22] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krahenbuhl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5390–5399, 2019. 2

[23] Rudolph E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82:35–45, 1960. 2

[24] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):140–153, 2018. 7

[25] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 4696–4704, 2015. 2

[26] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 2, 5

[27] Cheng-Hao Kuo and Ram Nevatia. How does person identity recognition help multi-person tracking? In *CVPR 2011*, pages 1217–1224, 2011. 2

[28] Tino Kutschbach, Erik Bochinski, Volker Eiselein, and Thomas Sikora. Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–5. IEEE, 2017. 7

[29] 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. 2

[30] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 6

[31] Yiming Liang and Yue Zhou. Lstm multiple object tracker combining multiple cues. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2351–2355, 2018. 2

[32] Chen Long, Ai Haizhou, Zhuang Zijie, and Shang Chong. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *ICME*, 2018. 6, 7

[33] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 2, 3, 7

[34] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 6, 7, 9, 10

[35] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 4055–4064, 2018. 3

[36] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, 2019. 3

[37] Nan Ran, Longteng Kong, Yunhong Wang, and Qingjie Liu. A robust multi-athlete tracking algorithm by exploiting discriminant features and long-term dependencies. In *International Conference on Multimedia Modeling*, 2019. 2

[38] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1

[39] Jimmy Ren, Xiaohao Chen, Jianbo Liu, Wenxiu Sun, Jiahao Pang, Qiong Yan, Yu-Wing Tai, and Li Xu. Accurate single stage detector using recurrent rolling convolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5420–5428, 2017. 1, 2

[40] 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*, volume 28, 2015. 1, 2, 6, 9, 10

[41] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 5

[42] Seyed H. Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3047–3055, 2015. 1

[43] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 6, 9

[44] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018. 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. 6

[46] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 6

[47] Fatemeh Saleh, Sadegh Aliakbarian, Hamid Rezatofighi, Mathieu Salzmann, and Stephen Gould. Probabilistic tracklet scoring and inpainting for multiple object tracking. *arXiv preprint arXiv:2012.02337*, 2020. 2

[48] Ricardo Sanchez-Matilla, Fabio Poiesi, and Andrea Cavallaro. Online multi-target tracking with strong and weak detections. In *European Conference on Computer Vision*, pages 84–99. Springer, 2016. 6, 7

[49] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 6

[50] Hao Sheng, Yang Zhang, Jiahui Chen, Zhang Xiong, and Jun Zhang. Heterogeneous association graph fusion for target association in multiple object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3269–3280, 2018. 2

[51] Julian Smith, Florian Particke, Markus Hiller, and Jörn Thielecke. Systematic analysis of the pmbm, phd, jpda and gnn multi-target tracking filters. In *2019 22th International Conference on Information Fusion*, pages 1–8, 2019. 1

[52] Roy L. Streit and Tod E. Luginbuhl. Maximum likelihood method for probabilistic multihypothesis tracking. In *Signal and Data Processing of Small Targets*, volume 2235, pages 394–405. International Society for Optics and Photonics, 1994. 1

[53] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2, 3

[54] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):104–119, 2019. 2

[55] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017. 2

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the*

31st International Conference on Neural Information Processing Systems, NIPS'17, pages 6000–6010, 2017. 2, 3, 4, 5, 9

[57] Xingyu Wan, Jinjun Wang, and Sanping Zhou. An online and flexible multi-object tracking framework using long short-term memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2

[58] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. In *European Conference on Computer Vision*, 2020. 2

[59] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[60] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 6

[61] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3988–3998, 2019. 6, 7

[62] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. How to train your deep multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6787–6796, 2020. 6, 7

[63] Bo Yang and Ram Nevatia. An online learned crf model for multi-target tracking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2034–2041, 2012. 2

[64] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2137, 2016. 1, 2, 6, 9, 10

[65] Junbo Yin, Wenguan Wang, Qinghao Meng, Ruigang Yang, and Jianbing Shen. A unified object motion and affinity model for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6768–6777, 2020. 6, 7

[66] Yang Zhang, Hao Sheng, Yubin Wu, Shuai Wang, Weifeng Lyu, Wei Ke, and Zhang Xiong. Long-term tracking with deep tracklet association. *IEEE Transactions on Image Processing*, 29:6694–6706, 2020. 2

[67] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, 2020. 2, 7

[68] Ji Zhu, Hua Yang, Nian Liu, Minyoung Kim, Wenjun Zhang, and Ming-Hsuan Yang. Online multi-object tracking with dual matching attention networks. In *European Computer Vision Conference*, 2018. 6, 7