Lost and Found: Overcoming Detector Failures in Online Multi-Object Tracking

Lorenzo Vaquero[1,2], Yihong Xu[3], Xavier Alameda-Pineda[4], Víctor M. Brea[2], and Manuel Mucientes[2]

[1] Fondazione Bruno Kessler, Italy
lvaquerootal@fbk.eu
[2] CiTIUS, Univ. of Santiago de Compostela, Spain
{victor.brea, manuel.mucientes}@usc.es
[3] Valeo.ai, France
yihong.xu@valeo.com
[4] Inria Grenoble, Univ. Grenoble Alpes, France
xavier.alameda-pineda@inria.fr

Abstract. Multi-object tracking (MOT) endeavors to precisely estimate the positions and identities of multiple objects over time. The prevailing approach, tracking-by-detection (TbD), first detects objects and then links detections, resulting in a simple yet effective method. However, contemporary detectors may occasionally miss some objects in certain frames, causing trackers to cease tracking prematurely. To tackle this issue, we propose BUSCA, meaning 'to search', a versatile framework compatible with any online TbD system, enhancing its ability to persistently track those objects missed by the detector, primarily due to occlusions. Remarkably, this is accomplished without modifying past tracking results or accessing future frames, i.e., in a fully online manner. BUSCA generates proposals based on neighboring tracks, motion, and learned tokens. Utilizing a decision Transformer that integrates multimodal visual and spatiotemporal information, it addresses the object-proposal association as a multi-choice question-answering task. BUSCA is trained independently of the underlying tracker, solely on synthetic data, without requiring fine-tuning. Through BUSCA, we showcase consistent performance enhancements across five different trackers and establish a new state-of-the-art baseline across three different benchmarks. Code available at: https://github.com/lorenzovaquero/BUSCA.

Keywords: 2D Tracking · Multi-target tracking · Online

1 Introduction

Multi-object tracking (MOT) entails the process of locating and identifying multiple objects over time within a scene. It is a crucial task in computer vision with applications spanning various domains such as robotics [15], autonomous vehicles [18, 62], and video surveillance systems [38]. The prevalent MOT paradigm

Fig. 1: Due to occlusions, detectors fail to locate many relevant elements on a scene (e.g., the woman in red). Accordingly, online multi-object trackers may lose track of some objects. With BUSCA, we propose a fully online framework that can be integrated into any online TbD tracker to persistently track those objects missed by the detector. Box colors represent object identities.

is tracking-by-detection (TbD) [8], where object trajectories are obtained by (i) first detecting objects and (ii) then associating detections. Although alternative frameworks have been proposed in the literature [1, 29], TbD has surfaced capitalizing on significant progress in object detection. Notably, over the past few years, center- [61, 72] and Transformer-based architectures [48, 61] have emerged. More recently, the MOT performance has been further improved thanks to the adoption of YOLO-based detectors [17, 39] coupled with a straightforward intersection-over-union (IoU) matching. This simple yet effective approach has even contributed to the renewed popularity of SORT [6, 13, 39].

Meanwhile, significant efforts in the community have been also dedicated to improving identity consistency within a trajectory. This is achieved by devising better association schemes [6, 13, 67, 73] or through re-identification (Re-ID) [40, 45]. However, these methods remain highly dependent on the availability of detections, which makes them susceptible to trajectory fragmentations.

Current state-of-the-art detectors are not perfect and fail to detect all the objects in a video. To have an idea, 17% of the detections in MOT17 [30] validation set are still missed by the YOLOX detector [17], and the extremely occluded objects (visibility = 0, provided in the ground-truth annotations) contribute 11.0 points to the MOTA score based on the standard MOT evaluation [10, 30]. Meanwhile, modern online trackers pause or terminate the tracking process during these situations where an object fails to be detected, leading to suboptimal results. We argue that more care should be taken in this regard, avoiding premature terminations of objects that genuinely exist. In this work, we introduce **BUSCA** (**B**uilding **U**nmatched trajectorie**S** **C**apitalizing on **A**ttention), which helps online TbD systems handle those objects, often highly occluded, overlooked

by the detector. BUSCA propagates unmatched tracks and, by design, can be applied to the outcome of any online TbD track assignment process.

Some works in the literature [11, 35, 44] focus on repairing fragmented tracks and improving trajectory continuity. However, these have so far been implemented through offline methods, as they alter decisions made on previous time steps (e.g., interpolating a trajectory after re-detection) and/or leverage future information. Thus, despite some of them claiming to be online, they should be considered as offline according to the widely accepted definition of 'online' in MOTChallenge [10, 30] where *"the solution has to be immediately available with each incoming frame and cannot be changed at any later time"*. The offline fashion makes them impractical for certain real-world applications and not comparable to online methods. Conversely, BUSCA is able to *persistently track undetected objects in a fully online setting*[5].

As an example illustrated in Fig. 1, some objects are missed due to low visibility even by a highly performant detector [17], causing the tracker to lose them. With BUSCA, we can enhance any TbD online tracker to continuously track those undetected objects without resorting to offline methods. To this end, BUSCA is built on a multi-choice question-answering Transformer that finds undetected objects given (i) *candidate* generated with a motion model (independent of the detector), (ii) *contextual information* derived from neighboring objects, and (iii) *previous observations* from the object of interest. These inputs are composed of visual and spatiotemporal information. The visual component characterizes object appearances while the spatiotemporal element encapsulates the size, center location, and timing of the object in a condensed format using an innovative spatiotemporal encoder.

In summary, the main contributions and novelties of this work are as follows:

- BUSCA is a *general* framework to persistently track those objects missed by the detector, in a fully online manner, without (i) modifying past tracking predictions (ii) or accessing future frames.
- BUSCA entails (i) a novel *Decision Transformer* inspired by multi-choice question-answering tasks, (ii) a *Proposal Generator* that relies on neighboring tracks, motion, and learned tokens, and (iii) an innovative *Spatiotemporal Encoder* that captures the size, location, and time of the objects. The network is trained independently from the underlying tracker and using synthetic data [14], without any fine-tuning on real MOT sequences.
- BUSCA can be seamlessly integrated on top of any online TbD tracker, as demonstrated in our comprehensive experiments where we systematically enhance the performance of five distinct trackers on standard benchmarks [10, 30], defining a new state-of-the-art among online trackers.

2 Related Work

End-to-end MOT methods model detection, tracking, and their implicit matching within a unified architecture. The most common approaches tackle this

[5] BUSCA strictly respects the 'online' definition, thus 'fully online'.

through identity embeddings [57], regression [1,53] or the recent use of attention mechanisms [5,16,29,66,73,74]. Nonetheless, this holistic design can create challenges during the joint training process [16] and, prevent these methods from being applicable to other trackers and leveraging leading-edge detectors. Consequently, these models have not yet superseded TbD techniques.

Tracking by detection (TbD) is an effective paradigm that decouples the MOT task into object detection and data association. This decomposition enables TbD methods [19,45,48,60,61,67,70,72] to benefit from classical [39,41,60], more advanced [23, 67] or self-constructed [48, 61, 72] detectors, coupled with diverse association processes such as hierarchical clustering [70], graph neural networks [19] or geometric cues [67].

In particular, center-based methods like CenterTrack [73] and TransCenter [61] alleviate the ambiguity in bounding boxes by predicting object center heatmaps in a CNN-based or Transformer-based architecture, respectively. Recently, ByteTrack [67] showcases remarkable results using a meticulously tuned YOLOX detector [17] paired with a simple IoU-based matching mechanism. This powerful detector has also revived SORT [3] with a stronger association mechanism in methods such as OC-SORT and StrongSORT [6,13]. Nevertheless, these TbD trackers remain highly vulnerable to missed detections. This issue motivates us to introduce BUSCA, a framework designed to improve any online TbD tracker by persistently tracking those objects overlooked by the detector.

Improving trajectory consistency, i.e., maintaining consistent object identities over time, is one of the main challenges of online multi-object trackers. Most of these methods rely on frame-by-frame association of detections solved via Hungarian matching [25]. However, pure motion-based associations [3, 4, 67] often encounter difficulties in crowded environments or moving-camera scenarios. As a result, other works turn to appearance-based techniques [24,37,40,46,51,58], hybrid cues [13, 26, 45, 50], or Transformer solvers [66, 73]. Notably, GHOST [45] redesigns the use of a ReID model and builds a simple yet strong baseline. In efforts to lessen the impact of occlusions, some methods aim to predict an object's visibility in order to adjust its detections' confidences [21] or re-weight the association matrix [69]. On the other hand, some strategies improve associations by hallucinating object trajectories [49] or by prompting re-detections in areas where occluders are present [27].

Nonetheless, unlike BUSCA, these more advanced association processes *remain heavily dependent on the detector as they operate on available detections.* [26] is a rare exception but at the cost of MOT performance drop.

Ensuring trajectory continuity is a non-trivial task that attempts to repair the trajectory of an object from the instant it is lost until it is re-identified again. Thus, most current trackers perform an extra *offline* post-processing step based on linear [67] or Gaussian-smoothed [13] interpolation. Some more sophisticated methods involve implementing a probabilistic model to retroactively insert missed detections [44], learning an additional Refind Module [35] to bridge these gaps, or 2D-to-3D lifting and performing motion forecasting in a bird's eye view [11]. Nevertheless, these strategies remain *offline* [10, 30] as they either al-

ter predictions on past time steps or take into account future frames, limiting their applicability in certain real-world scenarios. We introduce thus BUSCA, a framework that can be *built on top of any online TbD tracker* to enhance its continuity and consistency *in a fully online fashion*.

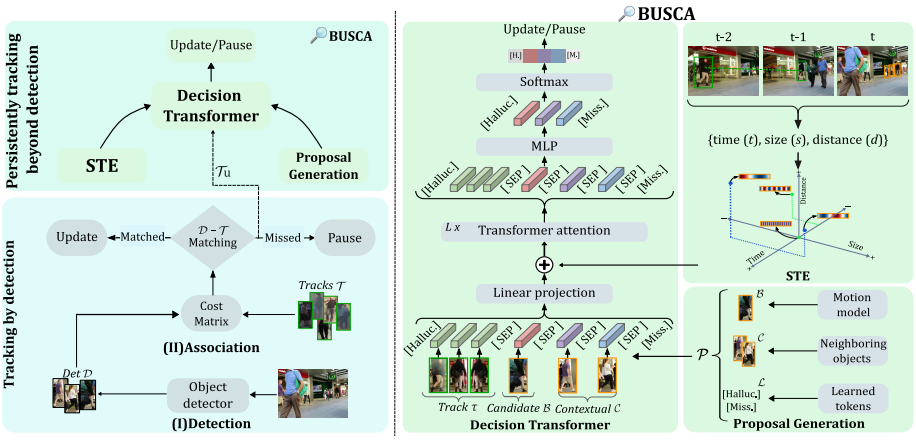


Fig. 2: The bottom-left panel depicts the tracking-by-detection (TbD) paradigm (Sec. 3), where a track is paused when the detector fails to locate the object. To address this issue, we integrate BUSCA into the online TbD tracker (Sec. 4) as shown in the top-left panel. This allows for the extension of trajectories of undetected objects by pairing them with proposals comprising *candidates* ($\mathcal{B}$), *contextual information* ($\mathcal{C}$) and *learned tokens* ($\mathcal{L}$) (Sec. 4.2) via an innovative decision Transformer (Sec. 4.1). Comprehensive details about the components of BUSCA are showcased in the right-hand panel. The *track* observations and proposals fed to the decision Transformer are made up of both appearance features (extracted with a convolutional backbone omitted here for clarity) and spatiotemporal cues for time, size, and distance encoded in a compact embedding through our novel spatiotemporal encoding (STE, Sec. 4.3).

3 TbD in a Nutshell

In the tracking by detection (TbD) paradigm, at a given frame a detector first produces a set $\mathcal{D} = \{\delta_1, ..., \delta_M\}$ of M detections, with each detection $\delta_i = \{a_i, c_i, \omega_i\}$ is defined by its appearance a_i (i.e., features of the image contained in the coordinates), coordinates c_i (object size and center location) and confidence score ω_i. These detections are used to propagate the position of a set $\mathcal{T} = \{\tau_1, ..., \tau_N\}$ of N active tracks, each represented by a time-ordered set $\tau_j = (o_{j,1}, \ldots, o_{j,Z})$ of observations $o_k = \{a_k, c_k\}$ over the past Z frames.

$\mathcal{D}$ is compared with $\mathcal{T}$, using coordinates and geometric cues [3, 67], appearance information [58], or both [45], yielding a cost matrix of size $N \times M$ whose optimal assignments are determined through Hungarian matching [25].

Thus, as shown in the bottom-left part of Fig. 2, correctly matched tracks are updated with the assigned detections, while those without a matching detection are paused. Having correct and sufficient detections for all tracks is critical, leading many trackers to resort to offline interpolation techniques to repair missing observations. In order to address this issue without resorting to offline interpolation, we present BUSCA, which tracks those undetected objects in a fully online fashion.

4 🔎 BUSCA: Finding Objects without Detections

Current detectors still fail to detect all the objects, especially in low-visibility situations i.e., heavy occlusions. Modern trackers heavily rely on the detection quality, thus naively stopping the tracking process whenever the detector fails. Therefore, BUSCA comes to help by saving those objects missed by the detector and finding where they are.

In particular, BUSCA is a fully online framework that can be coupled with any TbD tracker to persistently track those objects missed by the detector. As can be seen in the upper left part of Fig. 2, BUSCA receives unmatched tracks $\mathcal{T}_u$ and compares them with a set of proposals generated through a proposal generation process (Sec. 4.2). This comparison is carried out through a novel decision Transformer (Sec. 4.1), which uses an innovative spatiotemporal encoding (STE, Sec. 4.3) to aggregate information of different nature. This way, BUSCA can update the coordinates of those unmatched tracks or determine whether they have really left the scene.

4.1 Decision Transformer: To Be or Not To Be

Deciding whether to pause an undetected track or propagate its identity can be formulated as a multiple-choice question-answering task [36]. That is, given a question (the track τ) and a set of possible options (the proposals $\mathcal{P} = \{p_1, ..., p_J\}$, where $p_i = \{a_i, c_i\}$), the goal of the network is to find the correct answer (the decision of which proposal to match to the track) forming the assignment set $\mathcal{A} = \{\tau_j \mapsto p_i | \tau_j \in \mathcal{T}, p_i \in \mathcal{P}\}$. Inspired by this formulation, we propose to maintain undetected objects via a Transformer-based design that inputs different *proposals* and a *track*, outputting the best match, i.e., the proposal with the highest probability.

As shown on the right side of Fig. 2, our decision Transformer is implemented through an L-layer encoder model, which receives an input $\mathcal{I} = \{\tau, \mathcal{P}\}$, in which the past observations of the track are included. For each of the individual elements that make up the input (referred to as *tokens*), the appearance information a is processed by a convolutional backbone and projected to a lower dimensional space. This visual information of each token is then fused with its geometric cues c using our innovative spatiotemporal encoding (Sec. 4.3), to allow the Transformer to reason complex relationships between motion and visual features.

Within the decision Transformer, the input tokens are self-attended with each other, yielding refined tokens $\mathcal{J} = \{\overline{\tau}, \overline{\mathcal{P}}\}$ where the features most closely related to the track have been enhanced. Then, the elements of $\overline{\mathcal{P}}$ are fed to a shared-weight multi-layer perceptron (MLP) that generates one logit per token. After a Softmax operation, we output the probabilities that the track τ is assigned to each proposal p, allowing us to obtain $\mathcal{A}$ by finding the maximum probability. Finally, we update τ when it is successfully matched with a candidate proposal (See Sec. 4.2) or pause it otherwise. It should be noted that the MLP is share-weight, so as not to be restricted to any fixed input size.

4.2 Proposal Generation: Missing Puzzle Pieces

As with textual question-answering problems, the composition of the proposals $\mathcal{P}$ is one of the most critical aspects, and this is no different for our decision Transformer. $\mathcal{P} = \{\mathcal{B}, \mathcal{C}, \mathcal{L}\}$ is composed of candidates $\mathcal{B}$, contextual proposals $\mathcal{C}$, and learned proposals $\mathcal{L}$. As shown in the bottom-right of Fig. 2, $\mathcal{B}$ and $\mathcal{C}$ are extracted from the frame, while $\mathcal{L}$ is learned. BUSCA will keep a track τ active and update it with the proposal information if it is associated with any element from $\mathcal{B}$ and pause τ otherwise.

Generating the sets of proposals $\mathcal{B}$ and $\mathcal{C}$ is nontrivial given that none of the detections in $\mathcal{D}$ can be associated with τ. Given its reasonable performance [3, 13, 67], we opt for a simple yet effective Kalman filter [22] to predict a new observation of τ at the current frame. To this end, it is possible to obtain $\mathcal{B} = \{\text{Kalman}(\tau)\}$ without adding extra complexity to BUSCA, all while effectively managing complex motion scenarios, as evidenced in the supplementary material. Regarding the contextual proposals $\mathcal{C}$, their goal is to provide BUSCA with more information about the scene. $\mathcal{C}$ is composed of the Q closest observations within the neighborhood of τ, $V(\tau)$. Details for the computation of the maximum neighborhood distance for τ are given in the supplementary material.

The input proposals $\mathcal{P}$ of BUSCA also comprise a set $\mathcal{L} = \{$`[Halluc.]`, `[Miss.]`$\}$ of learned tokens that allow the Transformer to make complex decisions about the tracking process and pause τ if necessary. Specifically, `[Halluc.]` is learned to capture whether any observation o is corrupted (i.e., belonging to a different object) whereas `[Miss.]` handles if τ has left the scene or none of the elements of $\{\mathcal{B}, \mathcal{C}\}$ are suitable enough to be matched. Additionally, a separator token `[SEP]` borrowed from textual Transformers [36] is also learned to delimit each of the elements of $\mathcal{P}$.

4.3 Spatiotemporal Encoding (STE): Merging Modalities

Along with appearance features, spatiotemporal information is also crucial for making correct assignments. This information is however more complex to be encoded due to its multi-dimensionality (i.e., time-stamp t at which observations are recorded, the size s of the bounding box, and their distance d in the 2D coordinate space). To this end, we propose the spatiotemporal encoding (STE)

depicted on the top-right part of Fig. 2, which models these relationships between observations and allows its fusion with visual features so BUSCA can effectively learn complex relationships. Our spatiotemporal encoding supersedes the conventional positional encoding often implemented in Transformer models [52]. This encoding is generated through a two-step process comprising the *interplay mapping* and subsequent the *embedding projection*.

Interplay mapping. The encodings employed in visual Transformers rely on absolute values, which limit the network's overall adaptability and make them rely on interpolation techniques to handle diverse frame sizes [7,12,31]. Moreover, this method has consequential downsides for tracking tasks, as identical interactions might be represented differently depending on their specific occurrence (e.g. proximity between a track and an observation will be encoded differently depending on their absolute position within the frame or video).

To address this, our STE relies on a novel interplay mapping that models interactions relative to an anchor κ. In our specific use case, $\kappa = \{x_\kappa, y_\kappa, w_\kappa, h_\kappa, t_\kappa\}$ corresponds to the coordinates (i.e., object center, width, and height) and timestamp of the last known observation of the track $o \in \tau$. To this end, we can compute a spatiotemporal embedding $\{E^t, E^s, E^d\}$ comprising time, size, and distance, respectively, for each token $\iota \in \mathcal{I}$ as:

$$E^t = \sigma^t \left(t_\iota - t_\kappa\right) \tag{1}$$

$$E^s = \sigma^s \left(\log\left(\frac{w_\iota}{w_\kappa}\right) + \log\left(\frac{h_\iota}{h_\kappa}\right)\right) \tag{2}$$

$$E^d = \sigma^d \log\sqrt{\left(\frac{x_\iota - x_\kappa}{w_\kappa}\right)^2 + \left(\frac{y_\iota - y_\kappa}{h_\kappa}\right)^2} \tag{3}$$

where $\sigma^t, \sigma^s, \sigma^d$ are scaling factors. This relative representation boosts the generalization capacity of BUSCA and improves convergence during training.

Embedding Projection. After computing the interplay mapping between input tokens and τ, it is essential to make this representation compatible with both the transformer and the visual features. However, adding multiple independent sinusoidal functions could lead to potentially ambiguous information, according to [56]. To this end, it is necessary to establish a joint spatiotemporal encoding by expanding the function used in [52] to a 3-dimensional space. Given the Transformer's internal dimension of D^{Tr} channels, we equally distribute it among the three components of our spatiotemporal embedding $D = D^{\mathrm{Tr}}/3$. Therefore, for a given dimension E^Δ where $\Delta \in \{t, s, d\}$ we can compute its projected embedding PE^Δ:

$$PE^\Delta_{2i} = \sin\left(\frac{E^\Delta}{10000^{2i/D}}\right) \qquad PE^\Delta_{2i+1} = \cos\left(\frac{E^\Delta}{10000^{2i/D}}\right) \tag{4}$$

where $0 \leq i < D/2$. And subsequently concatenate the components of the different dimensions to create our compact spatiotemporal encoding $STE = \left(PE^t, PE^s, PE^d\right)$ for each one of the tokens $\iota \in \mathcal{I}$.

5 Experimental Results

In Sec. 5.1, we clarify the experimental settings along with the used datasets and metrics. In Sec. 5.2, we validate the necessity of BUSCA compared to the naive solutions and show that it can systematically extend tracks' lifespan, improving trajectory continuity without losing consistency. Subsequently, we empirically demonstrate the effectiveness of each component of BUSCA and justify its design choices. Once validated, we show in Sec. 5.3 that BUSCA is a plug-and-play component that consistently improves various trackers, setting new state-of-the-art performance in all tested benchmarks compared to other online methods. Finally, some successful and failure cases are qualitatively shown in Sec. 5.4.

5.1 Experimental Settings

We conduct our experiments on the widely-used MOT16 [30], MOT17 [30] and the crowded MOT20 [10] datasets. In contrast to other methods, we train BUSCA using solely synthetic data from MOTSynth [14], which consists of 764 full-HD videos recorded at 20 fps. For each training sample, we construct a track of length $Z = 11$ and randomly select 5 objects near τ to form a proposal set (current observation of τ is the positive candidate while objects with an overlap smaller than 0.5 are negatives. Additionally, we set a 15% probability of not sampling any positives (`[Miss.]` will be considered the correct option) and a 1% chance of altering observations within τ (`[Halluc.]` will be the correct option). Our training process focuses only on bounding box annotations and does not require any fine-tuning towards particular datasets or tracking systems. The computational cost of BUSCA is relatively small, with only 8.7M parameters and a runtime of 45ms per frame on a single NVIDIA RTX GPU (when integrated with [67], the whole system runs at roughly 13fps).

For the ablation, we focus on MOT17 with the widely-adopted split [45, 67, 72] that evenly divides each video sequence into training and validation sets. Unless otherwise stated, we employ ByteTrack [67] as our baseline tracker due to its state-of-the-art performance, but we remove its offline interpolation and its per-sequence curated thresholds. For the comparison with the state-of-the-art, we submit our test set results to the MOTChallenge servers and compare our approach with current *online* methods as defined in the challenge [10, 30].

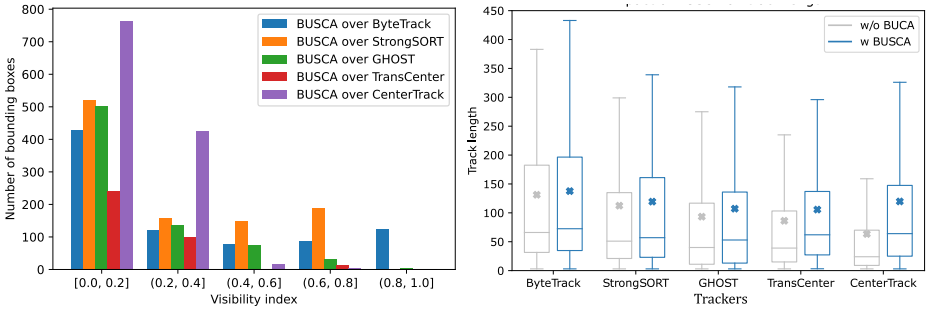
For evaluation, we report the standard metrics adopted by the MOTChallenge [9]. These include MOTA [2] reflecting the overall performance of a predicted trajectory; the recently introduced HOTA [28] that balances object coverage and identity preservation; IDF1 [43] focusing on association quality; IDentity SWitches (IDSW) to reflect identity consistency; and False Positives (FP) as well as False Negatives (FN) to assess detection performance. Additional experiments and implementation details can be found in the supplementary material.

5.2 Model Validation and Ablation

Naive approaches are not enough. Persistently tracking objects overlooked by the detector is not a trivial task and cannot be achieved with simpler naive ap-

Table 1: Comparison to different simpler solutions on MOT17 [30] val set. The difference with the baseline is depicted next to each metric. ByteTrack [67] is used as base tracker removing its offline interpolation and per-sequence thresholds, noted with ⋆.

	MOTA ↑	HOTA ↑	FN ↓	FP ↓
ByteTrack⋆	76.5	67.4	9120	**3410**
+ LD	75.3 (−1.2)	65.6 (−1.8)	8854 (**−266**)	4196 (+786)
+ IoU	75.4 (−1.1)	67.0 (−0.4)	**7588** (**−1532**)	5493 (+2083)
+ Mixed	76.6 (**+0.1**)	**67.6** (**+0.2**)	8393 (**−727**)	4063 (+653)
+ BUSCA (ours)	**77.1** (**+0.6**)	**67.6** (**+0.2**)	8326 (**−794**)	3889 (+479)



(a) Extra successfully rescued objects. **(b)** Impact of BUSCA on track length.

Fig. 3: (a) Analysis of the additional objects that BUSCA *successfully* locates when integrated with different trackers. The objects are grouped by their visibility [30]. (b) Analysis of the impact of BUSCA on the resulting track length in different trackers. Additional implementation details can be found in the supplementary material.

proaches. Specifically, ByteTrack [67] demonstrates that with a reliable detector, some low-score detections can be leveraged in a second-round association. One would then expect that **L**owering the **D**etection (**LD**) threshold $\epsilon = 0.01$ would provide further benefits during the tracks-detections matching. Another direct approach similar to BUSCA consists of using a motion model (e.g., Kalman filter) to estimate the track future coordinates and perform an extra round of associations based on motion and geometry cues like **IoU**. Alternatively, we also propose an extra recovery round based on **Mixed** cues (i.e. both IoU and appearance), as shown important for more robust associations [58].

As shown in Tab. 1, lower-score detections are not reliable and **+LD** increases FP (+786) with a slight decrease in FN (-266), leading to a MOTA (-1.2) and HOTA (-1.8) drop. This demonstrates that the leftover detections in [67] are not reliable and insufficient for finding lost objects and it is therefore necessary to leverage a motion model providing better candidates. However, not every candidate is reliable, and relying solely on **+IoU** associations does not improve MOT performance (-1.1/-0.4 in MOTA/HOTA). Adding visual cues with our **+Mixed** approach brings improvements, but the limited increase in MOTA (+0.1) evidences that this simple method still struggles to make correct assignments. Differently, **BUSCA** considers visual and spatiotemporal information

Table 2: Ablation on MOT17 [30] val set of the different components that comprise BUSCA. HLC=`[Halluc.]` learned token, MSS=`[Miss.]` learned token, STE=spatiotemporal encoding, CTX=contextual proposals. The difference with the baseline is depicted next to each metric. ByteTrack [67] is used as base tracker removing its offline interpolation and per-sequence thresholds.

Line	HLC	MSS	STE	CTX	**MOTA ↑**	**HOTA ↑**	**FN ↓**	**FP ↓**
1					76.5	67.4	9120	**3410**
2	✓				75.0 (−1.5)	66.3 (−1.1)	8395 (**−725**)	4911 (+1501)
3		✓			76.4 (−0.1)	67.3 (−0.1)	**8064** (**−1056**)	4513 (+1103)
4	✓	✓			76.5 (0.0)	67.1 (−0.3)	8656 (**−464**)	3853 (+443)
5	✓	✓	✓		76.7 (**+0.2**)	67.4 (0.0)	8528 (**−592**)	3851 (+441)
6	✓	✓		✓	76.9 (**+0.4**)	67.6 (**+0.2**)	8387 (**−733**)	3884 (+474)
7	✓	✓	✓	✓	**77.1** (**+0.6**)	**67.6** (**+0.2**)	8326 (**−794**)	3889 (+479)

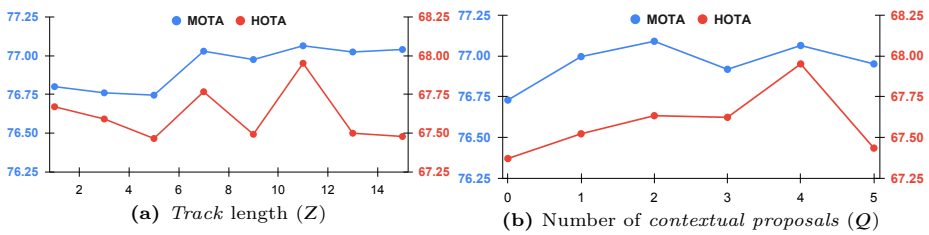
from the track, the candidate, and the context in a Transformer-based design, providing better decisions to prevent undetected tracks from being paused.

Longer trajectories with BUSCA. As illustrated in Fig. 3a, the efficacy of BUSCA is evident in its ability to *successfully* keep alive an extensive array of missing objects under different baselines. We observe that most of those saved objects have low visibility (i.e., under heavy occlusions), proving that BUSCA is particularly good at mitigating instances where the detector exhibits a proclivity for failure. Accordingly, BUSCA correctly extends the resulting track trajectories *in every tested tracker*, as demonstrated in Fig. 3b.

BUSCA component ablation. BUSCA relies on different components that ensure its proper operation and allow it to associate proposals and tracks accurately. In Table 2, we analyze the impact of the learned `[Miss.]` and `[Halluc.]` tokens, the spatiotemporal encoding, and the use of contextual information.

BUSCA may decide to pause a track either because it is a hallucinated track (`[Halluc.]` token), or because none of the candidates is suitable enough (`[Miss.]` token). Relying solely on the `[Halluc.]` token (Line 2) yields negative results, resulting in an additional +1501 false positives compared to the baseline. Conversely, if track termination is guided solely by the `[Miss.]` token (Line 3), the output remains marginally below the baseline with a decrease of −0.1 points in MOTA. The integration of these two learned tokens leads to improved performance (Line 4) because taking into account both conditions for whether to associate a track more accurately represents real-world situations.

By adding our spatiotemporal encoding *STE* (Line 5), the MOTA score is further increased by +0.2 points. Nonetheless, a high number of false negatives persist due to duplicated tracks occasionally kept alive. These tracks negatively impact the system when kept active, and so far BUSCA has had no way of identifying them. To address this issue, we integrate contextual proposals from nearby observations (Line 6), successfully reducing false negatives by −733 and resulting in a MOTA increase of +0.4 points. The best results are achieved when all components are integrated into BUSCA (Line 7).



(a) *Track* length (Z) **(b)** Number of *contextual proposals* (Q)

Fig. 4: Study of track length and number of contextual proposals used as input in our decision Transformer w.r.t. HOTA and MOTA performance.

Track length, contextual proposal size. Adhering to the definition of an online method, BUSCA considers the past observations of a track and its interaction with neighboring objects, learning deep relationships between motion and appearance. On Fig. 4a, we study the optimal amount of observations fed as input to BUSCA. The HOTA curve contains noisier observations, whereas MOTA displays an upward trend that starts to converge at $Z = 11$ where HOTA also achieves the best score. Regarding the maximum number of contextual proposals, from Fig. 4b, we observe that both curves have a positive slope which decays when $Q > 4$. We hypothesize this is due to the additional contextual proposals being too distant and uninformative on the track's environment.

5.3 State-of-the-Art Comparisons

By design, BUSCA can be seamlessly incorporated into any existing online TbD tracker. To illustrate its performance, we extensively integrate BUSCA into five diverse state-of-the-art trackers and compare them against the current state-of-the-art in online MOT. Our base trackers include the center-based CenterTrack [71] (CNN network) and TransCenter [61] (Transformer network); as well as the YOLOX-based ByteTrack [67] (IoU matching), StrongSORT [13] (appearance-enhanced association), and GHOST [45] (attentive Re-ID scheme). Evaluations were conducted on the test sets of MOT16 [30], MOT17 [30], and MOT20 [10]. As shown in Tab. 3, *BUSCA consistently improves the performance of all trackers in every benchmark for nearly all metrics, without requiring training on any real MOT data nor necessitating to be fine-tuned for any tracker.*

Remarkably, BUSCA *drastically* enhances both CenterTrack and TransCenter without the necessity for a recent state-of-the-art detector. For instance, in CenterTrack, we achieve a boost of +12 HOTA and +21 IDF1 in MOT20. Similarly, TransCenter also gets significantly improved due to a marked reduction in IDSW, thereby bolstering HOTA (e.g., +5.1/+8.6 in MOT17/20) and IDF1 (e.g., +8.6/+15 in MOT17/20). When paired with high-performing trackers such as ByteTrack and StrongSORT that rely on a potent YOLOX detector [17], BUSCA sets a new state-of-the-art for online multi-object tracking. Furthermore, BUSCA can also join efforts with identity-preserving methods like the advanced Re-ID mechanism in GHOST [45] to further enhance its performance.

Table 3: State-of-the-art comparison on MOT16, MOT17, and MOT20 test sets. ⋆ means that the offline interpolation and the per-sequence thresholds in ByteTrack [67] and OC-SORT [6] are removed for fair comparison. † and ‡ indicate reproduced results for GHOST [45] and StrongSORT [13] on MOT16 and for CenterTrack [72] on MOT20, respectively, due to their unavailability in the original works. Private detections are used. BUSCA consistently improves all baseline trackers in almost every metric, as shown in **bold**. Best results are highlighted in blue.

	MOT16				MOT17				MOT20			
	MOTA↑	HOTA↑	IDF1↑	IDSW↓	MOTA↑	HOTA↑	IDF1↑	IDSW↓	MOTA↑	HOTA↑	IDF1↑	IDSW↓
TubeTK [32]	66.9	50.8	62.2	1236	63.0	48.0	58.6	5727	–	–	–	–
CTracker [34]	67.6	48.8	57.2	1897	66.6	49.0	57.4	5529	–	–	–	–
QDTrack [33]	69.8	54.5	67.1	1097	68.7	53.9	66.3	3378	–	–	–	–
TraDeS [59]	70.1	53.2	64.7	1144	69.1	52.7	63.9	3555	–	–	–	–
MTrack [65]	72.9	–	74.3	642	72.1	–	73.5	2028	63.5	–	69.2	6031
MeMOT [5]	72.6	57.4	69.7	845	72.5	56.9	69.0	2724	63.7	54.1	66.1	1938
MeMOTR [16]	–	–	–	–	72.8	58.8	71.5	1902	–	–	–	–
GSDT [55]	74.5	56.6	68.1	1229	73.2	55.2	66.5	3891	67.1	53.6	67.5	3230
Decode-MOT [26]	74.7	60.2	73.0	1094	73.2	59.6	72.0	3363	67.2	54.5	69.0	2805
MOTR [66]	–	–	–	–	73.4	57.8	68.6	2439	–	–	–	–
OUTrack [27]	74.2	59.2	71.1	1328	73.5	58.7	70.2	4122	68.6	56.2	69.4	2223
FairMOT [68]	75.7	61.6	75.3	621	73.7	59.3	72.3	3303	61.8	54.6	67.3	5243
TrackFormer [29]	–	–	–	–	74.1	57.3	68.0	2829	68.6	54.7	65.7	1532
TransTrack [48]	–	–	–	–	74.5	-	63.9	3663	64.5	–	59.2	3565
AOH [21]	–	–	–	–	75.1	59.6	72.6	3312	67.9	55.1	70.0	2698
GTR [73]	–	–	–	–	75.3	59.1	71.5	2859	–	–	–	–
CrowdTrack [47]	–	–	–	–	75.6	60.3	73.6	2544	70.7	55.0	68.2	3198
OC-SORT⋆ [6]	–	–	–	–	76.0	61.7	76.2	2199	73.1	60.5	74.4	1307
SGT [20]	76.8	61.2	73.5	1276	76.3	60.6	72.4	4578	72.8	56.9	70.5	2649
CorrTracker [54]	76.6	61.0	74.3	1709	76.5	60.7	73.6	3369	65.2	–	69.1	5183
ReMOT [64]	76.9	60.1	73.2	742	77.0	59.7	72.0	2853	–	–	–	–
Unicorn [63]	–	–	–	–	77.2	61.7	75.5	5379	–	–	–	–
MTracker [69]	–	–	–	–	77.3	–	75.9	3255	66.3	–	67.7	2715
MO3TR-YOLOX [74]	–	–	–	–	77.6	60.3	72.9	2847	72.3	57.3	69.0	2200
CountingMOT [42]	77.6	62.0	75.2	1087	78.0	61.7	74.8	3453	70.2	57.0	72.4	2795
CenterTrack‡ [72]	69.6	–	60.7	2124	67.8	52.2	64.7	3039	45.8	31.8	36.6	6296
+ **BUSCA (ours)**	**70.4** (+0.8)	**55.7** (–)	**69.7** (+9.0)	**927** (-1197)	**68.9** (+1.1)	**55.1** (+2.9)	**68.8** (+4.1)	**2817** (-222)	**49.5** (+3.7)	**44.2** (+12)	**58.0** (+21)	**1370** (-4926)
TransCenter [61]	75.7	56.9	65.9	1717	76.2	56.6	65.5	5427	72.9	50.2	57.7	2625
+ **BUSCA (ours)**	**75.7** (+0.0)	**61.9** (+5.0)	**74.5** (+8.6)	**1038** (-679)	**76.2** (+0.0)	**61.7** (+5.1)	**74.1** (+8.6)	**3282** (-2145)	**73.9** (+1.0)	**58.8** (+8.6)	**72.4** (+15)	**1756** (-869)
GHOST† [45]	78.3	63.0	77.4	709	78.7	62.8	77.1	2325	73.7	61.2	75.2	1264
+ **BUSCA (ours)**	**78.5** (+0.2)	**63.2** (+0.2)	**77.5** (+0.1)	**707** (-2)	**79.0** (+0.3)	**62.9** (+0.1)	77.0 (-0.1)	**2295** (-30)	**74.2** (+0.5)	**61.3** (+0.1)	75.1 (-0.1)	**1251** (-13)
StrongSORT† [13]	78.3	63.8	78.9	437	78.3	63.5	78.5	1446	72.2	61.5	75.9	1066
+ **BUSCA (ours)**	**78.4** (+0.1)	**64.2** (+0.4)	**79.5** (+0.6)	**434** (-3)	**78.6** (+0.3)	**63.9** (+0.4)	**79.2** (+0.7)	**1428** (-18)	**72.7** (+0.5)	**61.8** (+0.3)	**76.3** (+0.4)	**1006** (-60)
ByteTrack⋆ [67]	78.2	62.8	77.2	892	78.9	62.8	77.1	2363	74.2	60.4	74.5	925
+ **BUSCA (ours)**	**78.5** (+0.3)	**63.3** (+0.5)	**77.9** (+0.7)	**743** (-145)	**79.3** (+0.4)	**63.1** (+0.3)	**77.7** (+0.6)	**2358** (-5)	**74.5** (+0.3)	**60.5** (+0.1)	74.4 (-0.1)	**920** (-5)

Lastly, recent tracking-by-attention methods [16,29,66,73,74] strive to create a fully end-to-end architecture that performs both object detection and track-detection matching within a single network. However, this streamlined process hinders their ability to easily incorporate new elements, such as a more powerful detector. This is illustrated by MOT3TR-YOLOX [74], a recent model that, despite adopting a more powerful YOLOX detection backbone, still underperforms TransCenter+BUSCA by −1.4 HOTA, −1.2 IDF1 in MOT17 and by −1.5 HOTA −3.4 IDF1 in MOT20. This underscores the superior performance of TbD methods and the opportunities that BUSCA brings, offering a plug-and-play module that systematically enhances state-of-the-art TbD trackers in a fully online manner without the need for retraining.

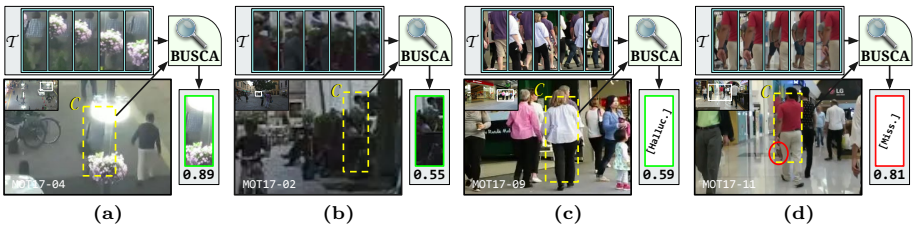


Fig. 5: Qualitative examples of BUSCA integrated into ByteTrack [67] for MOT17-val [9]. a, b, and c depict correct predictions while d shows a scenario where BUSCA incorrectly labels the pedestrian wearing a gray shirt as 'missing', even though the individual's left foot (highlighted with a red circle) remains visible. The values indicate the assignment confidence.

5.4 Qualitative Results

Fig. 5 showcases a series of qualitative visualizations. In Fig. 5a, the YOLOX detector [17] fails to detect the person obscured by the street lamp and flowers due to substantial occlusion. However, with BUSCA, we can successfully preserve his identity. A similar scenario unfolds in Fig. 5b, where the pedestrian in the background is accurately identified by BUSCA despite his minimal size and the scarce visibility of only his head. Fig. 5c illustrates a clearly spurious track created by ByteTrack [67] that does not correlate to any specific person. BUSCA correctly identifies it as a hallucination and deactivates it, effectively preventing any further false positives. Lastly, in Fig. 5d, due to the noisy track and the almost total occlusion, the pedestrian wearing a gray shirt is incorrectly labeled as missing, even though his left foot can still be spotted behind the man in red. Additional videos are provided in the supplementary material.

6 Conclusion

In this work, we present BUSCA, an innovative and plug-and-play framework that can enhance any online tracking-by-detection system to persistently track undetected objects in a fully online fashion. This implies that BUSCA *does not* alter the outputs of previous time steps or access future frames. To achieve this, our novel Decision Transformer associates tracks with proposals having both visual and spatiotemporal information, maintaining the identity of tracks in a lightweight manner and without any need for fine-tuning.

We extensively validate our proposed method with five distinct trackers, bringing systematic performance improvements and setting new state-of-the-art results across different benchmarks. For future work, we aim to factor in extreme motions via nonlinear multi-candidate proposals, incorporate 3D multimodal cues, and explore the use of BUSCA to override previous tracking decisions and fix incorrect associations. We hope that BUSCA can inspire future research towards fully online trackers without overly relying on the detectors.

Acknowledgements

This work was partially supported by the EU ISFP PRECRISIS (ISFP-2022-TFI-AG-PROTECT-02-101100539) project, the EU WIDERA PATTERN (HORIZON-WIDERA-2023-ACCESS-04-01-101159751) project, MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the Spanish Ministerio de Ciencia e Innovación (grant numbers PID2020-112623GB-I00 and PID2021-128009OB-C32). We thank Eloi Zablocki from Valeo.ai for the meaningful discussion.

References

1. Bergmann, P., Meinhardt, T., Leal-Taixé, L.: Tracking without bells and whistles. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 941–951 (2019)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F.T., Upcroft, B.: Simple online and realtime tracking. In: IEEE Int. Conf. Image Process. (ICIP). pp. 3464–3468 (2016)
4. Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS). pp. 1–6 (2017)
5. Cai, J., Xu, M., Li, W., Xiong, Y., Xia, W., Tu, Z., Soatto, S.: Memot: Multi-object tracking with memory. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 8080–8090 (2022)
6. Cao, J., Weng, X., Khirodkar, R., Pang, J., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 9686–9696 (2023)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 9630–9640 (2021)
8. Dai, Y., Hu, Z., Zhang, S., Liu, L.: A survey of detection-based video multi-object tracking. Displays **75**, 102317 (2022)
9. Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., Leal-Taixé, L.: MOTChallenge: A benchmark for single-camera multiple target tracking. Int. J. Comput. Vis. **129**(4), 845–881 (2021)
10. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I.D., Roth, S., Schindler, K., Leal-Taixé, L.: MOT20: A benchmark for multi object tracking in crowded scenes. CoRR **abs/2003.09003** (2020)
11. Dendorfer, P., Yugay, V., Osep, A., Leal-Taixé, L.: Quo vadis: Is trajectory forecasting the key towards long-term multi-object tracking? Adv. Neural Inf. Process. Syst. (NeurIPS) **35**, 15657–15671 (2022)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Repr. (ICLR) (2021)
13. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: Strongsort: Make deepsort great again. IEEE Trans. Multimedia (2023)

14. Fabbri, M., Brasó, G., Maugeri, G., Ošep, A., Gasparini, R., Cetintas, O., Calderara, S., Leal-Taixé, L., Cucchiara, R.: Motsynth: How can synthetic data help pedestrian detection and tracking? In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 10829–10839 (2021)
15. Gad, A., Basmaji, T., Yaghi, M., Alheeh, H., Alkhedher, M., Ghazal, M.: Multiple object tracking in robotic applications: Trends and challenges. Applied Sciences **12**(19) (2022)
16. Gao, R., Wang, L.: Memotr: Long-term memory-augmented transformer for multi-object tracking. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 9901–9910 (October 2023)
17. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021. CoRR **abs/2107.08430** (2021)
18. Guo, S., Wang, S., Yang, Z., Wang, L., Zhang, H., Guo, P., Gao, Y., Guo, J.: A review of deep learning-based visual multi-object tracking algorithms for autonomous driving. Applied Sciences **12**(21) (2022)
19. He, J., Huang, Z., Wang, N., Zhang, Z.: Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5299–5309 (2021)
20. Hyun, J., Kang, M., Wee, D., Yeung, D.: Detection recovery in online multi-object tracking with sparse graph tracker. In: IEEE Winter Conf. Appl. Comp. Vis. (WACV). pp. 4839–4848 (2023)
21. Jiang, M., Zhou, C., Kong, J.: AOH: online multiple object tracking with adaptive occlusion handling. IEEE Signal Process. Lett. **29**, 1644–1648 (2022)
22. Kalman, R.E.: A new approach to linear filtering and prediction theory. J. Fluids. Eng. **82**(1), 35–45 (1960)
23. Khan, A.H., Munir, M., van Elst, L., Dengel, A.: F2dnet: Fast focal detection network for pedestrian detection. In: IEEE Int. Conf. Pattern Recognit. (ICPR). pp. 4658–4664 (2022)
24. Kim, C., Li, F., Alotaibi, M., Rehg, J.M.: Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 9553–9562 (2021)
25. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics **52**(1-2), 83–97 (1955)
26. Lee, S.H., Park, D.H., Bae, S.H.: Decode-mot: How can we hurdle frames to go beyond tracking-by-detection? IEEE Trans. Image Process. **32**, 4378–4392 (2023)
27. Liu, Q., Chen, D., Chu, Q., Yuan, L., Liu, B., Zhang, L., Yu, N.: Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. Neurocomputing **483**, 333–347 (2022)
28. Luiten, J., Osep, A., Dendorfer, P., Torr, P.H.S., Geiger, A., Leal-Taixé, L., Leibe, B.: HOTA: A higher order metric for evaluating multi-object tracking. Int. J. Comput. Vis. **129**(2), 548–578 (2021)
29. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 8844–8854 (2022)
30. Milan, A., Leal-Taixé, L., Reid, I.D., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. CoRR **abs/1603.00831** (2016)
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2:

Learning robust visual features without supervision. CoRR **abs/2304.07193** (2023)
32. Pang, B., Li, Y., Zhang, Y., Li, M., Lu, C.: Tubetk: Adopting tubes to track multi-object in a one-step training model. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 6307–6317 (2020)
33. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 164–173 (2021)
34. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: European Conf. Comput. Vis. (ECCV). vol. 12349, pp. 145–161 (2020)
35. Qin, Z., Zhou, S., Wang, L., Duan, J., Hua, G., Tang, W.: Motiontrack: Learning robust short-term and long-term motions for multi-object tracking. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 17939–17948 (2023)
36. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. OpenAI Research pp. 1–12 (2018)
37. Rafi, U., Doering, A., Leibe, B., Gall, J.: Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In: European Conf. Comput. Vis. (ECCV). pp. 36–52 (2020)
38. Rani, J.U., Raviraj, P.: Real-time human detection for intelligent video surveillance: An empirical research and in-depth review of its applications. SN Comput. Sci. **4**(3), 258 (2023)
39. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 779–788 (2016)
40. Ren, H., Han, S., Ding, H., Zhang, Z., Wang, H., Wang, F.: Focus on details: Online multi-object tracking with diverse fine-grained representation. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11289–11298 (2023)
41. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
42. Ren, W., Chen, B., Shi, Y., Jiang, W., Liu, H.: Countingmot: Joint counting, detection and re-identification for multiple object tracking. CoRR **abs/2212.05861** (2022)
43. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conf. Comput. Vis. (ECCV). pp. 17–35 (2016)
44. Saleh, F.S., Aliakbarian, S., Rezatofighi, H., Salzmann, M., Gould, S.: Probabilistic tracklet scoring and inpainting for multiple object tracking. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 14329–14339 (2021)
45. Seidenschwarz, J., Brasó, G., Elezi, I., Leal-Taixé, L.: Simple cues lead to a strong multi-object tracker. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 13813–13823 (2023)
46. Shuai, B., Berneshawi, A.G., Li, X., Modolo, D., Tighe, J.: Siammot: Siamese multi-object tracking. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 12372–12382 (2021)
47. Stadler, D., Beyerer, J.: On the performance of crowd-specific detectors in multi-pedestrian tracking. In: IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS). pp. 1–12 (2021)

48. Sun, P., Jiang, Y., Zhang, R., Xie, E., Cao, J., Hu, X., Kong, T., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple-object tracking with transformer. CoRR **abs/2012.15460** (2020)
49. Tokmakov, P., Li, J., Burgard, W., Gaidon, A.: Learning to track with object permanence. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 10840–10849 (2021)
50. Vaquero, L., Brea, V.M., Mucientes, M.: Real-time siamese multiple object tracker with enhanced proposals. Pattern Recognit. **135**, 109141 (2023)
51. Vaquero, L., Mucientes, M., Brea, V.M.: Tracking more than 100 arbitrary objects at 25 fps through deep learning. Pattern Recognit. **121**, 108205 (2022)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Adv. Neural Inf. Process. Syst. (NeurIPS). pp. 5998–6008 (2017)
53. Wan, X., Cao, J., Zhou, S., Wang, J., Zheng, N.: Tracking beyond detection: Learning a global response map for end-to-end multi-object tracking. IEEE Trans. Image Process. **30**, 8222–8235 (2021)
54. Wang, Q., Zheng, Y., Pan, P., Xu, Y.: Multiple object tracking with correlation learning. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3876–3886 (2021)
55. Wang, Y., Kitani, K., Weng, X.: Joint object detection and multi-object tracking with graph neural networks. In: IEEE Int. Conf. Rob. Autom. (ICRA). pp. 13708–13715 (2021)
56. Wang, Z., Liu, J.: Translating math formula images to latex sequences using deep neural networks with sequence-level training. Int. J. Document Anal. Recognit. **24**(1), 63–75 (2021)
57. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: European Conf. Comput. Vis. (ECCV). vol. 12356, pp. 107–122 (2020)
58. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: IEEE Int. Conf. Image Process. (ICIP). pp. 3645–3649 (2017)
59. Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: An online multi-object tracker. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 12352–12361 (2021)
60. Xiang, Y., Alahi, A., Savarese, S.: Learning to track: Online multi-object tracking by decision making. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 4705–4713 (2015)
61. Xu, Y., Ban, Y., Delorme, G., Gan, C., Rus, D., Alameda-Pineda, X.: Transcenter: Transformers with dense representations for multiple-object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 7820–7835 (2023)
62. Xu, Y., Chambon, L., Chen, M., Alahi, A., Cord, M., Perez, P., et al.: Towards motion forecasting with real-world perception inputs: Are end-to-end approaches competitive? In: IEEE Int. Conf. Rob. Autom. (ICRA) (2024)
63. Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., Lu, H.: Towards grand unification of object tracking. In: European Conf. Comput. Vis. (ECCV). vol. 13681, pp. 733–751 (2022)
64. Yang, F., Chang, X., Sakti, S., Wu, Y., Nakamura, S.: Remot: A model-agnostic refinement for multiple object tracking. Image Vis. Comput. **106**, 104091 (2021)
65. Yu, E., Li, Z., Han, S.: Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 8824–8833 (2022)

66. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: MOTR: end-to-end multiple-object tracking with transformer. In: European Conf. Comput. Vis. (ECCV). vol. 13687, pp. 659–675 (2022)
67. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European Conf. Comput. Vis. (ECCV). pp. 1–21 (2022)
68. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vis. **129**(11), 3069–3087 (2021)
69. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Robust multi-object tracking by marginal inference. In: European Conf. Comput. Vis. (ECCV). vol. 13682, pp. 22–40 (2022)
70. Zhao, K., Imaseki, T., Mouri, H., Suzuki, E., Matsukawa, T.: From certain to uncertain: Toward optimal solution for offline multiple object tracking. In: IEEE Int. Conf. Pattern Recognit. (ICPR). pp. 2506–2513 (2020)
71. Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., Tao, D.: Transvod: End-to-end video object detection with spatial-temporal transformers. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 7853–7869 (2023)
72. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conf. Comput. Vis. (ECCV). vol. 12349, pp. 474–490 (2020)
73. Zhou, X., Yin, T., Koltun, V., Krähenbühl, P.: Global tracking transformers. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 8761–8770 (2022)
74. Zhu, T., Hiller, M., Ehsanpour, M., Ma, R., Drummond, T., Reid, I., Rezatofighi, H.: Looking beyond two frames: End-to-end multi-object tracking using spatial and temporal transformers. IEEE Trans. Pattern Anal. Mach. Intell. (2022)