No Train Yet Gain:
Towards Generic Multi-Object Tracking in Sports and Beyond

Tomasz Stanczyk[1,2] Seongro Yoon[1,2] Francois Bremond[1,2]
[1] Inria, France [2] Université Côte d'Azur, France
tomasz.stanczyk@inria.fr, seong-ro.yoon@inria.fr, francois.bremond@inria.fr

Abstract

Multi-object tracking (MOT) is essential for sports analytics, enabling performance evaluation and tactical insights. However, tracking in sports is challenging due to fast movements, occlusions, and camera shifts. Traditional tracking-by-detection methods require extensive tuning, while segmentation-based approaches struggle with track processing. We propose McByte, a tracking-by-detection framework that integrates temporally propagated segmentation mask as an association cue to improve robustness without per-video tuning. Unlike many existing methods, McByte does not require training, relying solely on pre-trained models and object detectors commonly used in the community. Evaluated on SportsMOT, DanceTrack, SoccerNet-tracking 2022 and MOT17, McByte demonstrates strong performance across sports and general pedestrian tracking. Our results highlight the benefits of mask propagation for a more adaptable and generalizable MOT approach. Code will be made available at https://github.com/tstanczyk95/McByte.

1. Introduction

Multi-object tracking (MOT) is to involve tracking objects across video frames while maintaining consistent object IDs. It gets initiated to detect objects in each frame and associate them across consecutive frames. MOT can be applied to tracking players and performers in various sport settings to facilitate both team and individual performance analysis and statistics. Nevertheless, these sport settings are still challenging due to dynamic movement of the targets, blurry objects and occlusions, posing strong demands to tracking solutions.

Tracking-by-detection [4, 13, 25, 33, 34, 38], one of the intuitive approaches, firstly detects the target objects with bounding boxes in each frame and associates them with those from previous frames, based on cues such as position, appearance, and motion. Then, the resulting matches form "tracklets" over consecutive frames. However, these methods often require extensive hyper-parameter tuning for each dataset or even per single sequence, reducing their generalizability and limiting their application across different datasets and various sport scenarios.

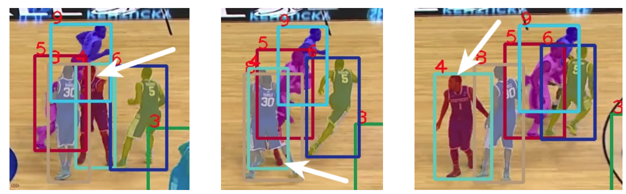


Figure 1. Mask propagation module can be helpful in cases of severe occlusion. The person with the red mask is tracked only by its limited visible parts (pointed by white arrows), particularly in the middle image. Input image data from [10].

Segmentation mask-based methods [7, 28], on the other hand, generate masks to cover objects and track them across video frames. Trained on large datasets, these methods aim to capture the semantics of image patches, making them more generic. However, they are not designed for MOT, lacking robust management for tracking multiple entities and struggling to detect new objects entering the scene, especially team sport players with dynamically moving camera. Additionally, these methods rely entirely on mask predictions for object positioning, which can be problematic when the predictions are noisy or inaccurate.

In this paper, we explore applying temporal mask propagation method as an association cue to assess its effectiveness in MOT of challenging scenarios in sport media. We propose a novel tracking-by-detection method that combines mask propagation module and detections to improve the association. Since the applied temporal mask propagation model has been trained on a large dataset, it makes the entire tracking process more generic. Unlike existing tracking-by-detection methods, our approach does not require sensitive tuning of hyper-parameters for each dataset or video sequence, i.e. it is not parametric.

A combination of a linear motion prediction model such as Kalman filter [18] and detection matching based on intersection over union (IoU), which is a baseline for most of the tracking-by-detection methods [4, 13, 25, 33, 34], is not suitable on its own for settings with highly dynamic motion and blurry objects, in tracking the sport players or performers. We believe that the temporally propagated mask might be able to handle this. However, providing segmentation mask ground truths for each video frame, also for blurry objects, is a huge burden. In our approach, we do not perform any costly training of the mask segmentation or temporal propagator. Instead, we apply well-studied models pre-trained on a huge corpus of data without specifically tuning them on MOT datasets, but with carefully incorporating them in our method.

Tracking multiple objects at once often involves handling challenging occlusions. To be specific, when many players try to take over the ball, where only a small part of the subject might be visible. Temporally propagated mask can be especially helpful in such cases, when the visible shape can considerably differ from the subject. A visual example is presented in Fig. 1, where person with ID 4 is severely occluded, but the mask (in red) still identifies their shoe and helps to maintain tracking the person.

We note explicitly that incorporating temporally propagated mask, which also involves temporal coherency, is different than using a static mask coming directly from an image segmentation model [19] independently per each frame. Using the mask temporal propagation as an association cue within the problem of MOT has not been done before.

We evaluate our incorporation of the temporally propagated mask as an association cue against a baseline tracker. We show clear benefits for MOT, including (but not limited to) sport settings, by handling challenging situations such as ambiguous occlusions and blurry tracked objects. Our tracker is tested on four MOT datasets, while for a fair comparison, we use the same, pre-trained object detectors used in the community. Our method outperforms tracking-by-detection algorithms on SportsMOT [10], DanceTrack [31], SoccerNet-tracking 2022 [9] and MOT17 [26]. These results highlight the advantages of using mask propagation, eliminating the need for per-sequence hyper-parameter tuning. Solely with Kalman filter, provided detection, mask segmentation and propagation models, we obtain meaningful results and performance on four different datasets.

Our contribution in this work is summarized as follows: (i) We design a tracking-by-detection algorithm, based on Kalman filter and a pre-trained mask propagation model, which does not require any expensive training (*no train, yet gain*). (ii) We propose a practical idea adapting a temporally propagated object segmentation mask as an effective association cue incorporated into an MOT tracking algorithm. The tracker overcomes the limitations of mask-based approaches by performing proper tracklet management and including other important association cues as well as the limitations of the baseline tracking-by-detection approaches, by making the tracking process more robust and generic. (iii) We propose practical policies enforcing regulated usage of the propagated mask in a controlled manner, which help handle common situations in sport settings, such as occlusions and blurry objects.

2. Related Work

Multi-object tracking approaches. Transformer-based approaches [14, 15, 36, 39] use attention mechanisms for end-to-end learning of tracking trajectories and object associations but require extensive training data. Other MOT methods include global optimization (offline tracking) [5] and joint detection-tracking approaches [35, 37].

Tracking-by-detection methods detect objects in each frame and associate them into tracklets. ByteTrack [38], a strong baseline, uses the YOLOX [16] detector and Kalman Filter [18] to associate tracklets via intersection over union (IoU). Several methods extend ByteTrack: OC-SORT [4] improves state estimation with virtual trajectories during occlusion, StrongSORT [13] incorporates re-identification (re-ID) features, camera motion compensation, and NSA Kalman filter [12], C-BIoU [33] enlarges bounding boxes for better association, and HybridSORT [34] adds confidence modeling and height-modulated IoU.

Although effective on MOT datasets like MOT17 [26], these algorithms are highly sensitive to parameters. ByteTrack, for example, performs per-sequence tuning of detection thresholds in its practical implementation, and its extensions [4, 13, 25, 34] also rely on extensive parameter adjustments. This process is costly and impractical for large datasets like SportsMOT [10], DanceTrack [31], and SoccerNet-tracking [9], as well as for generalizing across different sports. In contrast, our method leverages temporally propagated segmentation masks as an association cue, eliminating the need for per-sequence tuning and improving robustness across diverse settings.

Mask Temporal Propagation. XMem [6], based on the Atkinson-Shiffrin Memory Model [1], enables long-term segmentation mask tracking in video object segmentation (VOS). Its successor, Cutie [8], enhances segmentation by incorporating object encoding from mask memory for better background differentiation. While image segmentation models like SAM [19] generate initial masks at specific frames, mask temporal propagation models infer and update masks across subsequent frames.

Although effective in their domains, XMem and Cutie are not directly suited for MOT as they do not involve bounding boxes and can produce inaccurate mask predictions [6, 8]. To address this, we propose a novel MOT algorithm that integrates temporally propagated masks with bounding boxes, enhancing tracking performance.

Mask-Based Tracking Systems. Segmentation mask mod-

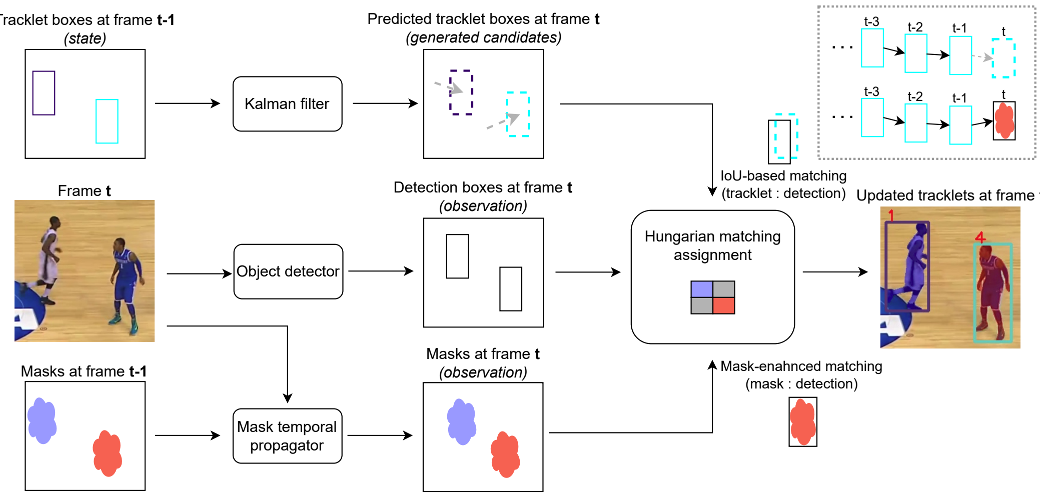


Figure 2. The overview of our proposed tracking pipeline. **Top:** Existing tracklet bounding box positions are considered as the tracker state from the previous frame, t-1. Using Kalman filter, candidate tracklet positions are generated and considered as a predicted next state. **Middle:** Current frame t is passed to object detector. Then, initial association cost matrix is computed based on IoU between the bounding boxes of tracklet candidate positions and detections. **Bottom:** Mask temporal propagator takes the current frame t as an input and produces the masks, which are then considered as additional observation. The cost matrix is enriched with the information based on the matching between masks and detections. Finally, the cost matrix is passed to the Hungarian matching algorithm minimizing the association cost and the tracklets at frame t are updated with the matched detections.

els have been applied to tracking. DEVA [7], an extension of XMem [6], introduces decoupled video segmentation and bi-directional propagation, involving masks and bounding boxes. Grounded SAM 2 combines Grounding Dino [22] and SAM 2 [28] for bounding box tracking and object ID maintenance. MASA [21], a SAM-based [19] mask feature adapter [3, 17], supports video segmentation and object tracking by matching detections across frames. However, these mask-based approaches lack robust tracklet management and struggle with occlusions, new object entries, and missed detections.

3. Proposed method

3.1. Preliminaries

In tracking-by-detection methods [4, 13, 25, 33, 34, 38], detection bounding boxes of the same objects are joined over the frames and form so called tracklets. In the current frame, new detections are associated with the existing tracklets from the previous frames. This process uses association cues such as object position, motion and displacement information to build a cost matrix reflecting a cost of potential match for each tracklet-detection pair. The association problem is then considered as the bipartite matching problem and solved with the Hungarian matching algorithm [20] to minimize the overall matching cost. Pairs with costs above the pre-defined matching threshold are excluded. Matched detections extend the existing tracklets.

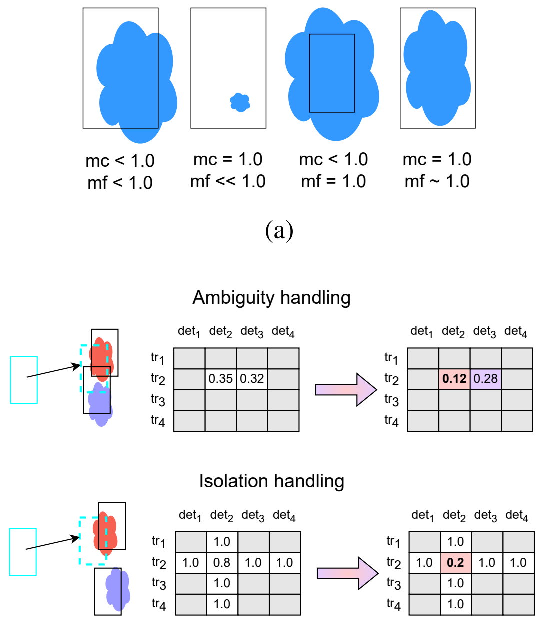
In our baseline, ByteTrack [38], new detections are split into high and low confidence groups, handled separately in the association process. The baseline uses intersection over union (IoU) as the primary association metric. Tracklet next state position bounding boxes are predicted using a linear motion model, Kalman Filter [18], and compared to the newly observed detection bounding boxes using IoU scores. Cost matrix entries are filled with 1-IoU for each tracklet-detection pair. Tracklet bounding boxes are updated with the matched detection bounding boxes. Unmatched detections are used to initiate new tracklets, while tracklets unmatched for too long are terminated. For more details, we refer the reader to the baseline paper [38].

In our work, we study a temporally propagated segmentation mask as a powerful association cue for MOT. We combine the mask and bounding box information to create our novel **m**asked-**c**ued algorithm, which we call McByte. Fig. 2 shows the overview of our tracking pipeline with inclusion of the propagated mask as an association cue.

3.2. Mask creation and handling

We design the following approach of mask handling to synchronise it with the processed tracklets. Each tracklet gets its own mask, which is then propagated across frames to keep it up-to-date. We use an image segmentation model [19] to create a mask for each new tracklet. It is performed only for a newly appeared object and to initialize a new mask.

Separately, during the next frames, a mask propagator model [8] is used to update the object masks. The masks



(a)

(b)

Figure 3. (a) Cases showing the differences in mc and mf values of a temporally propagated mask (in blue) within a bounding box. The most optimal case for the mask to provide a good guidance is the last one, where both mc and mf are as close to 1 as possible. (b) Ambiguity and isolation handling with the mask as a cue. Ambiguity occurs when the IoU-based costs are low and similar for more than one entry in a row (or column) of the cost matrix. Isolation occurs when relevant cost matrix entries contain too high values, not allowing for the association, and at the same time there is no ambiguity.

are used in the tracklet-detection association process, as detailed in Sec. 3.3. We process the propagated masks in sync with the tracklet lifespan, creating new masks for new tracklets and removing them when a tracklet is terminated.

3.3. Regulated use of the mask

Mask propagator might sometimes return erroneous masks, as seen in related works [6, 8]. Therefore, it is essential to regulate the mask use as a guiding association cue.

In the association process, we update the cost matrix entries using the temporally propagated (TP) mask in two particular, separate cases, for which we refer to as *ambiguity* and *isolation*. In cases of ambiguity, a tracklet could match multiple detections or vice versa. Ambiguity often arises from IoU-based matches when tracked objects are close, causing significant overlap in bounding boxes and thus similar IoU. If IoU-based cost is below the matching threshold for more than one tracklet-detection pair, we treat it as ambiguity.

In cases of isolation, a detection could potentially match a tracklet without any ambiguity, but their bounding boxes are too far from each other and the overlap is too small. The IoU-based cost is too high (above the matching threshold) to be considered for a match. Isolation might happen when the tracked objects are blurred or when camera movement is abrupt and the next state tracklet bounding boxes cannot be matched with observed detection bounding boxes.

For each potential ambiguous or isolated tracklet-detection match, we apply a strategy consisting of the following conditions. *(1) We check if the considered tracklet's TP mask is actually visible on the scene.* Subjects can be entirely occluded resulting in no mask prediction at the current frame. *(2) We check if the mask returned by the propagator is confident enough.* We average the confidence probability of all pixels of the mask assigned to a tracklet and check if it is above the set mask confidence threshold. Further, we compute two key ratios between the TP mask and a detection bounding box, measured with the number of pixels:

- the bounding box coverage of the mask, denoted as mc:

$$mc^{i,j} = \frac{|mask(tracklet_i) \cap bbox_j|}{|mask(tracklet_i)|} \tag{1}$$

- the mask fill ratio of the bounding box, denoted as mf:

$$mf^{i,j} = \frac{|mask(tracklet_i) \cap bbox_j|}{|bbox_j|} \tag{2}$$

where $mask(\cdot)$ denotes the TP mask assigned to the tracklet and $|\cdot|$ denotes the cardinality of the set. Note that all $mc, mf \in [0, 1]$. In Fig. 3(a), we show how mc and mf can vary depending on TP mask and detection bounding box position. After that, we consider two more conditions. *(3) We check if the mask fill ratio of the bounding box mf is sufficiently high.* Very low values can indicate TP masks which are noisy or come from another, wrong tracklet. However, we allow low values reflecting only the visible parts of a tracked object. *(4) We check if the bounding box coverage of the mask mc is sufficiently high.* Bounding boxes can sometimes be inaccurate, so we allow this value to be slightly below 1.0. Only if all the four conditions hold, we update the corresponding tracklet-detection pair entry in the IoU-based cost matrix:

$$costs^{i,j} = \begin{cases} costs^{i,j}_{IoU} - mf^{i,j}, & \text{if cond. } \textit{(1)-(4)} \text{ satisfied.} \\ costs^{i,j}_{IoU}, & \text{otherwise.} \end{cases} \tag{3}$$

where $costs^{i,j}$ denotes the final association cost between the pair of tracklet i and detection j, and $costs^{i,j}_{IoU}$ denotes the original IoU-based cost for this pair. We apply this process to each entry of the cost matrix, representing each possible tracklet-detection match pair. In Fig. 3(b), we show how TP mask as an association cue influences the cost matrix and guides the association. With this fusion of the available information, we consider both modalities, TP masks and bounding boxes to enhance the association process. In cases of ambiguity or isolation, TP mask assigned to a tracklet can guide the association with a new and suitable detection. The updated cost matrix, enriched by the mask cue, is passed to the Hungarian matching algorithm to find optimal tracklet-detection pairs.

Following conditions *(1)-(4)* ensures that the TP mask cue is controlled, and the cost matrix is updated only when the mask is reliable. Further, using mc directly to influence the cost matrix could be misleading, as multiple TP masks could fully fit within the same bounding box, all resulting in $mc = 1.0$. Therefore, we use mc only as a gating condition and mf to influence the cost matrix.

Our baseline is optimized for bounding boxes, so we retain the use of the Hungarian matching algorithm over the cost matrix, but we carefully incorporate the TP mask cue to enhance the association process.

3.4. Handling camera motion issues

When the camera moves, tracklet and detection bounding boxes may become less accurate. To address this, we also integrate camera motion compensation (CMC) into our process to enhance the accuracy of bounding box estimates. Our approach follows the existing methods [13, 25]. Specifically, we compute a warp (transformation) matrix that accounts for camera movement, based on extracted image features, and apply this matrix to the next state predicted tracklet bounding boxes. This helps adjust for the camera motion, making the tracklet predictions from the Kalman Filter [18] and the associations with detections more accurate, improving the overall tracking performance. For key-point extraction, we use the ORB (Oriented FAST and Rotated BRIEF) approach [30].

4. Experiments and discussion

4.1. Implementation details

For object detections, we follow the baseline [38] and use the YOLOX [16] detector pre-trained on the relevant dataset, unless stated otherwise. Detections are split into high and low confidence sets based on a threshold. Unlike the baseline, which adjusts this per sequence, we use a fixed 0.6 threshold across all sequences, matching the baseline's default when no per-sequence tuning is applied.

For the thresholds of mask confidence, mc and mf (Sec. 3.3) we set the values of 0.6, 0.9 and 0.05 respectively. We set these values considered that they must be high enough depending on their definition. We fix the same values for all sequences and all datasets. Changing these parameters around these values does not affect much performance of McByte making them not sensitive and generic.

For mask creation, we use SAM[19] (ViT_b model, original weights) for fair comparison with related works.

Method	HOTA	IDF1	MOTA
basline [38]: no mask	47.1	51.9	88.2
$a1$: either mask or no assoc.	48.6	44.4	80.8
$a2$: either mask or IoU for assoc.	45.3	41.5	82.2
$a3$: IoU and mask if ambig. or isol.	56.6	57.0	89.5
$a4$: $a3$ + mask confidence	57.3	57.7	89.6
$a5$: $a4$ + mf	58.8	60.1	89.6
$a6$: $a5$ + mc	62.1	63.4	89.7
McByte: $a6$ + cmc	**62.3**	**64.0**	**89.8**

Table 1. Ablation study on DanceTrack [31] validation set listing the effects of the imposed constraints on using the temporally propagated mask as an association cue.

Method	HOTA	IDF1	MOTA
Baseline, SportsMOT val	69.0	77.9	97.5
McByte, SportsMOT val	**83.9**	**83.6**	**98.9**
Baseline, DanceTrack val	47.1	51.9	88.2
McByte, DanceTrack val	**62.3**	**64.0**	**89.8**
Baseline, SoccerNet-tracking 2022 test	72.1	75.3	94.5
McByte, SoccerNet-tracking 2022 test	**85.0**	**79.9**	**96.8**
Baseline, MOT17 val	68.4	80.2	78.2
McByte, MOT17 val	**69.9**	**82.8**	**78.5**

Table 2. Ablation study comparing McByte with the baseline [38] on four different datasets. As SoccerNet-tracking 2022 [9] does not contain validation set split, we report the results on the test set.

Cutie[8] (Cutie base mega weights) is used as the mask temporal propagator.

4.2. Datasets and evaluation metrics

We evaluate McByte on four person tracking datasets, primarily in sports settings, while also testing on an additional dataset to assess generalizability. Results are presented on SportsMOT[10], DanceTrack[31], SoccerNet-tracking 2022[9], and MOT17[26], using dataset-specific detection sources per community standards for fair comparison.

SportsMOT [10] features basketball, volleyball, and football scenes with diverse indoor and outdoor court views. It includes fast camera movement and variable player speeds. We use YOLOX pre-trained on SportsMOT for detections.

DanceTrack [31] includes dancers with highly non-linear motion and subtle camera movements, while the number of individuals remains mostly constant. We use YOLOX pre-trained on DanceTrack for detections.

SoccerNet-tracking 2022 [9] features soccer match videos with fast-moving players who appear similar within teams. Camera movement is constant, and oracle detections are provided.

MOT17 [26] involves tracking pedestrians in public spaces under varying lighting, density, and camera stability. We use YOLOX pre-trained on MOT17 and report results as it is a widely used benchmark in the MOT community.

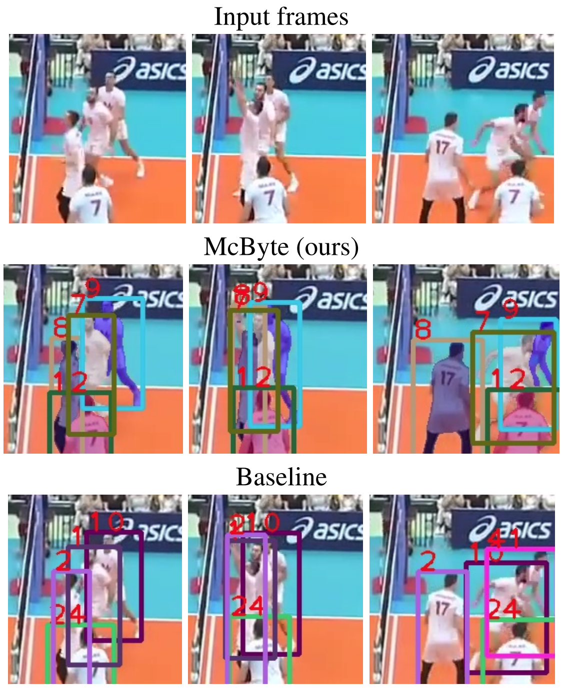


Figure 4. Example comparison with baseline [38] in a challenging volleyball setting. McByte can handle the association of ambiguous sets of bounding boxes. Players IDs 7 and 8 are well maintained despite the temporary jam. ID 9 is also well kept. In case of baseline, the corresponding player IDs are not maintained: After the jam, the player with ID 1 changes their ID to 10 and the player who previously had ID 10, now gets a new ID 41. ID 1 is lost.

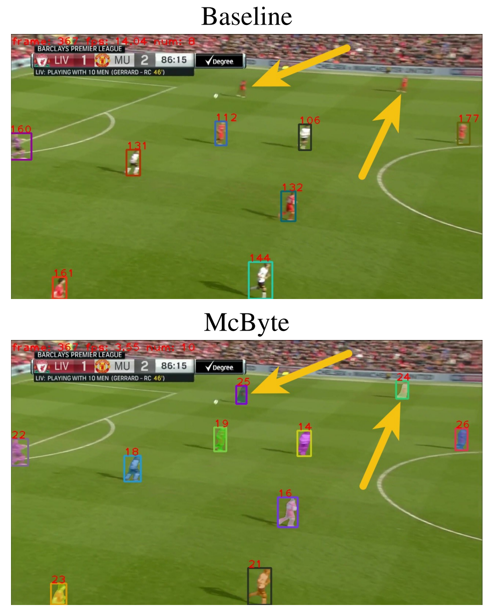


Figure 5. Example comparison with baseline [38] in a challenging football setting. McByte can maintain the tracklets of the blurry players (pointed by yellow arrows) caused by the abrupt camera movement.

We report three standard MOT metrics: HOTA[23], IDF1[29], and MOTA [2], focusing on HOTA (association, detection, and localization) and IDF1 (tracking quality and identity preservation). MOTA, which primarily evaluates detection quality, is included for completeness. Higher values indicate better performance.

We remark that we do not train object detectors, and for all datasets we use the same detections as baseline and other compared methods.

4.3. Ablation studies

We perform an ablation study to demonstrate the impact of incorporating the temporally propagated (TP) mask as an association cue along with the conditions discussed in Sec. 3.3. We evaluate the following variants:

- $a1$: Uses only the TP mask signal for association if the mask is visible for the given tracklet, without ambiguity/isolation checks. The value of $1 - mf^{i,j}$ is directly assigned to $costs^{i,j}$ in Eq. (3). No association occurs if there is no mask.
- $a2$: Similar to $a1$, but if the TP mask is unavailable, intersection over union (IoU) scores are used for association.
- $a3$: Adds an ambiguity/isolation check. If the TP mask is visible, mask and bounding box information are fused as shown in Eq. (3). If no mask is available, initial IoU scores are used.
- $a4$: Builds on $a3$, incorporating the mask confidence check (condition *(2)*).
- $a5$: Extends $a4$ by adding the mf value check from condition *(3)*.
- $a6$: Further extends $a5$ with the mc value check from condition *(4)*.

The results of each variant are listed in Tab. 1. In variant $a1$, where only the TP mask signal is used for association, we can see that despite HOTA increase, IDF1 decreases with respect to the baseline. It is caused by the fact that the mask use is uncontrolled and chaotic. With TP mask possibly providing incorrect results, the association cues can be misleading. If we perform the association either based only on TP mask or only on IoU (depending on the availability of the mask), as in variant $a2$, we might face an inconsistency of the cues between tracklets and detections from the same frame and the next frames. This might lead to performance degradation. However, when we use properly both cues fusing the available information (variant $a3$), we can observe significant performance gain. We explain it as the algorithm is initially designed to work on bounding boxes while TP mask is a valuable guiding cue which can improve existing association mechanisms. When the mask signal is the only cue or not properly fused with the IoU cue, a lower performance might be obtained (as in $a1$ and $a2$).

Adding the conditional check based on TP mask confi-

Method	HOTA	IDF1	MOTA
ByteTrack [38]	64.1	71.4	95.9
MixSort-Byte [10]	65.7	74.1	96.2
OC-SORT [4]	73.7	74.0	96.5
MixSort-OC [10]	74.1	74.4	96.5
GeneralTrack [27]	74.1	76.4	69.8
DiffMOT [24]	76.2	76.1	97.1
McByte (ours)	**76.9**	**77.5**	**97.2**

Table 3. Comparing McByte with state-of-the-art tracking-by-detection algorithms on SportsMOT test set [10].

Method	HOTA	IDF1	MOTA
ByteTrack [38]	47.7	53.9	89.6
MixSort-Byte [10]	46.7	53.0	85.5
OC-SORT [4]	55.1	54.9	92.2
Deep OC-SORT [25]	61.3	61.5	92.3
StrongSORT++ [13]	55.6	55.2	91.1
Hybrid-SORT [34]	65.7	67.4	91.8
GeneralTrack[27]	59.2	59.7	91.8
DiffMOT [24]	63.4	64.0	92.7
McByte (ours)	**67.1**	**68.1**	**92.9**

Table 4. Comparing McByte with state-of-the-art tracking-by-detection algorithms on DanceTrack test set [31].

Method	HOTA	IDF1	MOTA
ByteTrack [38]	72.1	75.3	94.5
OC-SORT [4]	82.0	76.3	**98.3**
McByte (ours)	**85.0**	**79.9**	96.8

Table 5. Comparing McByte with state-of-the-art tracking-by-detection algorithms on SoccerNet-tracking 2022 test set [9].

Method	HOTA	IDF1	MOTA
With parameter tuning per sequence			
ByteTrack [38]	63.1	77.3	80.3
MixSort-Byte [10]	64.0	78.7	79.3
StrongSORT++ [13]	64.4	79.5	79.6
OC-SORT [4]	63.2	77.5	78.0
MixSort-OC [10]	63.4	77.8	78.9
Deep OC-SORT [25]	64.9	80.6	79.4
Hybrid-SORT [34]	64.0	78.7	79.9
Without parameter tuning per sequence			
ByteTrack [5]	62.8	77.1	78.9
GeneralTrack [27]	64.0	78.3	**80.6**
DiffMOT [24]	**64.2**	79.3	79.8
McByte (ours)	**64.2**	**79.4**	80.2

Table 6. Comparing McByte with state-of-the-art tracking-by-detection algorithms on MOT17 test set [26].

dence (variant $a4$) further improves the performance, because sometimes mask might be uncertain or incorrect providing misleading association guidance. Adding the minimal mf value check (variant $a5$) also provides performance gain, because this check filters out the tracklet TP masks which could be considered as a noise or tiny parts of people almost entirely occluded. Another performance gain can be observed with the minimal mc value check (variant $a6$). This check determines if the TP mask of the tracked person is actually within the bounding box and not too much outside it. Since the detection box might not be perfect, we allow small parts of the tracklet mask to be outside the bounding box, but its major part must be within the bounding box, so that the TP mask can be used for guiding the association between the considered tracklet-detection pair. As it is shown, it further helps. Finally, we add the camera motion compensation (CMC), denoted as McByte in Tab. 1, which also provides some performance gain.

We compare our McByte tracking algorithm to the baseline [38] across the four datasets, as shown in Tab. 2. Performance gain is significant as a whole since improvement is always present, in all cases examined. DanceTrack [31] features mostly continuous non-linear motion and many substantial occlussions, where the TP mask signal is particularly helpful in tracking the subjects, whereas IoU might struggle. SportsMOT [10] and SoccerNet-tracking 2022 [9] feature similar outfits among the players, more abrupt motion, blurry objects, e.g. due to fast camera movement after the ball, and occlusions caused by the nature of the team sports. TP mask can handle these situations very well as shown in Figs. 4 and 5, and further improves the performance. Full frame images of Fig. 4 and more visual examples are available in Appendix C due to the space limits.

Although MOT17 [26] is not sports-specific, it is widely used in the MOT community, so we evaluate McByte on it. As shown in Tab. 2, McByte achieves competitive performance and consistently improves tracking scores, demonstrating its robustness and general applicability.

4.4. Comparison with state of the art tracking-by-detection methods

We compare McByte with state-of-the-art tracking-by-detection algorithms across the test sets of the four diverse datasets. The results in Tabs. 3 to 6 show that McByte outperforms the other methods on HOTA, IDF1 and MOTA on SportsMOT [10] (Tab. 3) and DanceTrack [31] (Tab. 4). In case of SoccerNet-tracking 2022 [9] (Tab. 5) and MOT17 [26] (Tab. 6), McByte outperforms the other methods on HOTA and IDF1, and achieves the second best MOTA score. We note that MOTA metrics mostly reflect the detection quality of the tracklets. Our proposed method works very well on datasets including settings with different sports, i.e. basketball, volleyball, football and dance. Further, it also achieves satisfactory results on another dataset involving pedestrians with less dynamic and more linear movement which, we believe, demonstrates its generalizability and wide applicability. While we do not perform any costly training or tuning on the evalu-

	SoccerNet-tracking 2022			MOT17			DanceTrack			SportsMOT		
Method	HOTA	IDF1	MOTA	HOTA	IDF1	MOTA	HOTA	IDF1	MOTA	HOTA	IDF1	MOTA
McByte (ours)	85.0	79.9	96.8	64.2	79.4	80.2	67.1	68.1	92.9	76.9	77.5	97.2
OC-SORT [4]	82.0	76.3	98.3	63.2	77.5	78.0	55.1	54.9	92.9	73.7	74.0	96.5
C-BIoU [33]	89.2	86.1	99.4	64.1	79.7	81.1	60.6	61.6	91.6	-	-	-
C-BIoU impl.	72.2	76.4	95.4	62.4	77.1	79.5	45.8	52.0	88.4	-	-	-

Table 7. Comparison with the tracking-by-detection methods which do not require training and use the same detections (Sec. 4.5). Comparing with C-BIoU is not direct as it involves manual tuning of important hyper-parameters. Since C-BIoU does not provide the code, we implement it following all the information in the paper (*"C-BIoU impl."*).

ated datasets, our method still reaches surpassing results on HOTA and IDF1 and competitve results on MOTA.

On MOT17 [26], unlike the baseline [38] and its extensions [4, 13, 25, 34], we do not tune parameters per sequence, aiming for a generalizable tracker. Among non-tuned trackers, McByte achieves the best scores. We also include the result of ByteTrack [38] not being tuned per sequence as reported in [5] (*"ByteTrack [5]"*). Compared to tuned methods, McByte improves over the baseline [38] and remains competitive with other approaches.

The other types of tracking methods, such as transformer-based [14, 15, 36, 39], global optimization [5] and joint detection and tracking [35, 37] require a lot of training data and might use other detections, which is not the focus of this work. For reference, we list their results together with other approaches and ours in Appendix B.

4.5. Comparison with non-trainable methods

In Tab. 7, we compare McByte with other non-trainable tracking-by-detection algorithms using the same detections. McByte, OC-SORT [4], and C-BIoU [33] achieve strong performance. McByte outperforms OC-SORT on all datasets, though OC-SORT achieves higher MOTA (reflecting the tracklet detection accuracy) on SoccerNet-tracking 2022, where oracle detections are provided.

C-BIoU [33] is not directly comparable, as it requires human-dependent tuning of buffer scales, critical hyper-parameters for each dataset. Values are provided for DanceTrack [31], while a separate technical report [32] lists different settings for SoccerNet-tracking [9]. Unfortunately, C-BIoU's code is not provided. However, the authors kindly share technical details, allowing us to implement it with fixed hyper-parameters across all datasets ("C-BIoU impl."). Differences in the performance reflect the sensitivity to the hyper-parameter tuning.

4.6. Comparison with other methods using mask

We evaluate mask-based tracking methods DEVA [7], Grounded SAM 2 [22, 28], and MASA [21] on the SportsMOT validation set. Each method is tested with its original settings and with YOLOX [16] trained on SportsMOT, as used in our baseline [38] and McByte for fair comparison. Results in Tab. 8 show that McByte outperforms all other mask-based methods.

Method	HOTA	IDF1	MOTA
DEVA, original settings	39.3	37.3	-109.6
DEVA, with YOLOX	42.4	42.1	-57.0
Grounded SAM 2, original settings	45.9	44.7	-13.9
Grounded SAM 2, with YOLOX	66.1	70.2	91.4
MASA, original settings	39.4	35.7	-27.2
MASA, with YOLOX	73.6	71.2	97.0
McByte (ours)	**83.9**	**83.6**	**98.9**

Table 8. Comparison with the other tracking methods using segmentation mask: DEVA [7], Grounded SAM 2 [19, 22] and MASA [21] on SportsMOT validation set [10]. We compare the variants with original settings and with the same object detector (YOLOX) as in McByte.

DEVA[7] lacks a tracklet management system, leading to negative MOTA scores when errors exceed object counts[11]. Adding YOLOX improves performance by refining object selection but remains limited. Grounded SAM 2[22, 28] uses segment-based tracking, though merging segments into tracklets can be inconsistent. MASA[21] has difficulty handling longer occlusions, occasionally missing detections, which affects MOT performance.

These results highlight that existing mask-based methods are unsuitable for MOT, particularly in sports. In contrast, McByte effectively combines temporally propagated mask-based association with bounding box processing and tracklet management, making it better suited for MOT tasks in sports and beyond.

More results and details on DEVA, Grounded SAM 2, and MASA experiments, also on other datasets (DanceTrack [31], MOT17 [26]), are available in Appendix A.

5. Conclusion

We introduce a temporally propagated segmentation mask as an association cue for MOT, focusing on sports tracking. Our approach incorporates mask propagation into tracking-by-detection by fusing mask and bounding box information while following our practical policies to enhance the performance. McByte requires no training or tuning, relying only on pre-trained models and object detectors for fair comparison. Results across four datasets highlight its effectiveness in sports and its generalizability to other scenarios.

Acknowledgement

This work has been supported by the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011014370 made by GENCI.

References

[1] R.C. Atkinson and R.M. Shiffrin. Human memory: A proposed system and its control processes. pages 89–195. Academic Press, 1968. 2

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 2008. 6

[3] Deblina Bhattacharjee, Sabine Süsstrunk, and Mathieu Salzmann. Vision transformer adapters for generalizable multitask learning, 2023. 3

[4] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9686–9696, 2023. 1, 2, 3, 7, 8

[5] Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. Unifying short and long-term tracking with graph hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22877–22887, 2023. 2, 7, 8

[6] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 4

[7] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 3, 8

[8] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4, 5

[9] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3502, 2022. 2, 5, 7, 8

[10] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 5, 7, 8

[11] Patrick Dendorfer, Aljossa Ossep, Anton Milan, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixe. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129:1–37, 2021. 8

[12] Yunhao Du, Junfeng Wan, Yanyun Zhao, Binyu Zhang, Zhihang Tong, and Junhao Dong. Giaotracker: A comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2809–2819, 2021. 2

[13] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 1, 2, 3, 5, 7, 8

[14] Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9901–9910, 2023. 2, 8

[15] Ruopeng Gao, Yijun Zhang, and Limin Wang. Multiple object tracking as id prediction, 2024. 2, 8

[16] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2, 5, 8

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[18] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. 1, 2, 3, 5

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 5, 8

[20] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 3

[21] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. *CVPR*, 2024. 3, 8

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 8

[23] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, pages 1–31, 2020. 6

[24] Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based

multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19321–19330, 2024. 7

[25] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813*, 2023. 1, 2, 3, 5, 7, 8

[26] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, 2016. arXiv: 1603.00831. 2, 5, 7, 8

[27] Zheng Qin, Le Wang, Sanping Zhou, Panpan Fu, Gang Hua, and Wei Tang. Towards generalizable multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19004, 2024. 7

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 3, 8

[29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision – ECCV 2016 Workshops*, pages 17–35, Cham, 2016. Springer International Publishing. 6

[30] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: an efficient alternative to sift or surf. pages 2564–2571, 2011. 5

[31] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5, 7, 8

[32] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. The second-place solution for cvpr 2022 soccernet tracking challenge. *arXiv preprint arXiv:2211.13481*, 2022. 8

[33] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4799–4808, 2023. 1, 2, 3, 8

[34] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-sort: Weak cues matter for online multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6504–6512, 2024. 1, 2, 3, 7, 8

[35] En Yu, Zhuoling Li, Shoudong Han, and Hongwei Wang. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*, 25:2686–2697, 2022. 2, 8

[36] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 8

[37] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 2, 8

[38] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. 2022. 1, 2, 3, 5, 6, 7, 8

[39] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22056–22065, 2023. 2, 8