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

L. Vaquero[1,2], Y. Xu[3], X. Alameda-Pineda[4], V. Brea[2], and M. 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. This keynote paper summarizes BUSCA, a work published in the European Conference on Computer Vision 2024. 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, 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 multi-modal 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. The prevalent MOT paradigm is tracking-by-detection (TbD), where object trajectories are obtained by (i) first detecting objects and (ii) then associating detections.

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 [3] validation

* Original paper: Lost and Found: Overcoming Detector Failures in Online Multi-Object Tracking. European Conf. Comp. Vision, 2024. ECCV is ICORE A* conf.

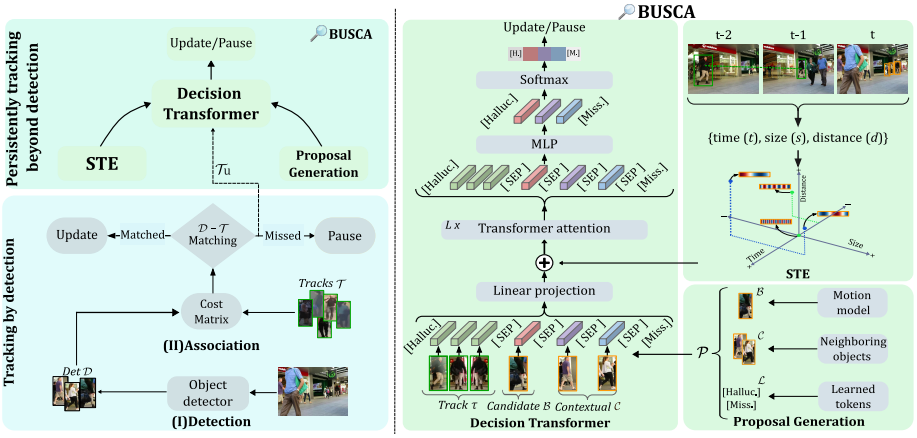


Fig. 1. The bottom-left panel depicts the tracking-by-detection (TbD) paradigm, where a track is paused when the detector fails to locate the object. To address this issue, we integrate BUSCA into the online TbD tracker as shown in the top-left panel. This allows for the extension of trajectories of undetected objects by pairing them with proposals comprising *candidates* ($\mathcal{B}$), *contextual information* ($\mathcal{C}$) and *learned tokens* ($\mathcal{L}$) via an innovative decision Transformer. 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).

set are still missed by the YOLOX detector [2], 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 [3,4]. 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.

2 BUSCA: Finding Objects without Detections

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. 1, BUSCA receives unmatched tracks $\mathcal{T}_u$ and compares them with a set of proposals generated through a proposal generation process. This comparison is carried out through a novel decision Transformer, which uses an innovative spatiotemporal encoding (STE) 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.

Deciding whether to pause an undetected track or propagate its identity can be formulated as a multiple-choice question-answering task [5]. 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. 1, 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, 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 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.

3 Experimental Results

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 [9] (CNN network) and TransCenter [7] (Transformer network); as well as the YOLOX-based ByteTrack [8] (IoU matching), StrongSORT [1] (appearance-enhanced association), and GHOST [6] (attentive Re-ID scheme). Evaluations were conducted on the test sets of MOT16 [3], MOT17 [3], and MOT20 [4]. As shown in table 1, 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.

Acknowledgments. This work was partially supported by the EU ISFP PRECRISIS (ISFP-2022-TFI-AG-PROTECT-02-101100539) project, the EU WIDERA PATTERN (HORI-ZON-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

Table 1. State-of-the-art comparison on MOT16, MOT17, and MOT20 test sets. ⋆ means that the offline interpolation and the per-sequence thresholds in ByteTrack [8] are removed for fair comparison. † and ‡ indicate reproduced results for GHOST [6] and StrongSORT [1] on MOT16 and for CenterTrack [10] 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↓
CenterTrack‡ [10]	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 [7]	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† [6]	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† [1]	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⋆ [8]	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)

from Valeo.ai for the meaningful discussion. **The authors have no competing interests to declare.**

References

1. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: Strongsort: Make deepsort great again. IEEE Trans. Multimedia (2023)
2. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021. CoRR **abs/2107.08430** (2021)
3. Milan, A., Leal-Taixé, L., Reid, I.D., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. CoRR **abs/1603.00831** (2016)
4. P. Dendorfer et al.: MOT20: A benchmark for multi object tracking in crowded scenes. CoRR **abs/2003.09003** (2020)
5. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. OpenAI Research pp. 1–12 (2018)
6. 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) (2023)
7. 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)
8. Y. Zhang et al.: Bytetrack: Multi-object tracking by associating every detection box. In: European Conf. Comput. Vis. (ECCV). pp. 1–21 (2022)
9. 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)
10. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conf. Comput. Vis. (ECCV). vol. 12349, pp. 474–490 (2020)