Benchmarking Multi-Object Tracking on Deep-Sea Video Footage Using HOTA

Elaine Liu, University of Virginia

Mentors: Lonny Lundsten

Summer 2026

Keywords: Multi-Object Tracking; Benchmark; Machine Learning

1. Abstract

Benchmarking multi-object tracking (MOT) model performance is an essential step in model development, as it enables data scientists to evaluate model detection and tracker performance on human generated 'test' data, allowing researchers to make fair, consistent comparisons between models and trackers, and, also, optimize model/tracker performance. In this study, a benchmark video dataset was developed to assess MBARI's 452k object detection model alongside several trackers. The benchmark dataset consists of four video sequences representing midwater and benthic deep-sea environments. Model performance was evaluated using HOTA (Higher Order Tracking Accuracy), a metric that balances detection, localization, and association accuracy. Results indicate that differences in the benthic and midwater environment did not have a significant effect on detection and tracking performance. However, fine-tuning tracker hyperparameters can have a significant, positive impact on overall detection/tracking performance.

2. Introduction

As the volume of video footage being captured continues to grow, so does the demand for efficient processing and analysis of data. MBARI stores their digitized video archive on a networked, multi-petabyte, solid state drive system which is managed by a database registry called MBARI Media Management (M3).Video observations are entered and archived in a searchable, SQL Server database called Video Annotation and Reference System, a database holding over 11 million observations from 37 years of archived deep-sea footage (Schlining & Stout, 2006). MBARI's Video Lab is tasked with annotating these videos, or in other words, staff identify and classify animals, geologic features, and equipment deployments and recoveries viewed in this footage. As a reference, a 1 hour video can take up to 2-3 hours to annotate. Given the large and increasing volumes of data to be annotated, using machine learning techniques to automate this process can significantly reduce time, effort, and cost. Being able to efficiently process these videos helps us better understand the deep sea environment in real-time, with the goal of potentially using them for real-time inferencing on future camera deployments.

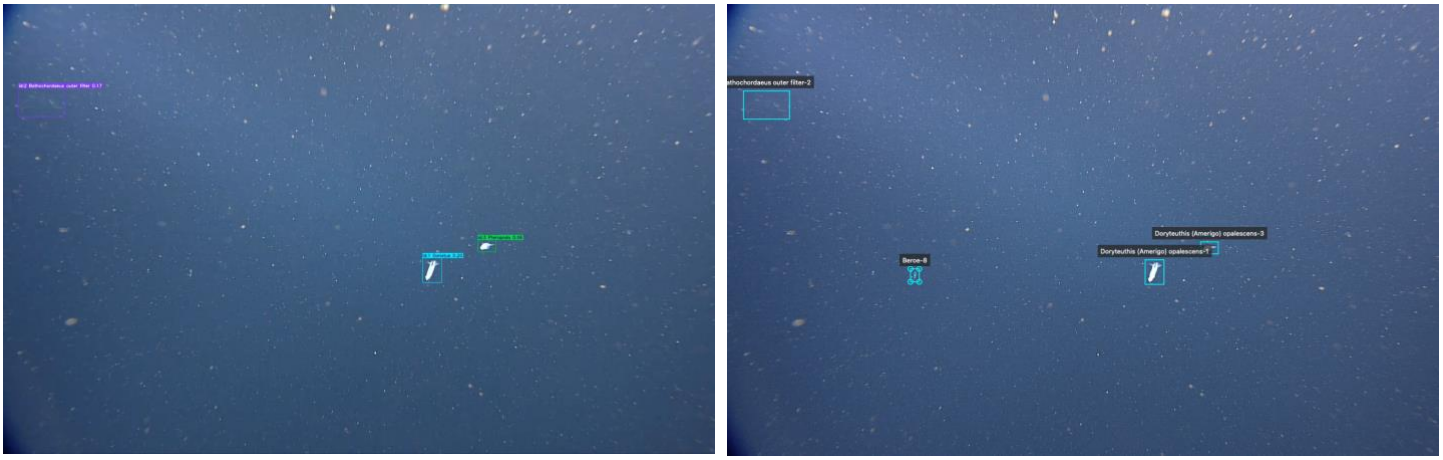
Multi-object tracking is made up of 3 main components: detection, localization, and association. Detection refers to what the object is, localization is where the object is, and association is the object's trajectory throughout the video. The object detection model handles detection and localization by generating bounding boxes in every single video frame. The tracker is responsible for association, and does so by linking detections together, from one frame to the next, correlating these detected objects. Thus, to successfully perform multi-object tracking, both the model and tracker need to be working accurately and cohesively.

To assess the accuracy of the model and tracker, a benchmark dataset is needed as a baseline for evaluation. Many benchmarks used in multi-object tracking to date are largely focused on tracking pedestrian and vehicle movement for the development of autonomous vehicles (Zhang et al., 2023). However, multi-animal tracking presents a unique challenge: individuals of the same species have uniform appearances, making it much more difficult for trackers to maintain consistent IDs across frames (Zhang et al., 2023). AnimalTrack is the first benchmark dedicated to multi-animal tracking and includes a variety of animals such as chickens, geese, dolphins, etc.

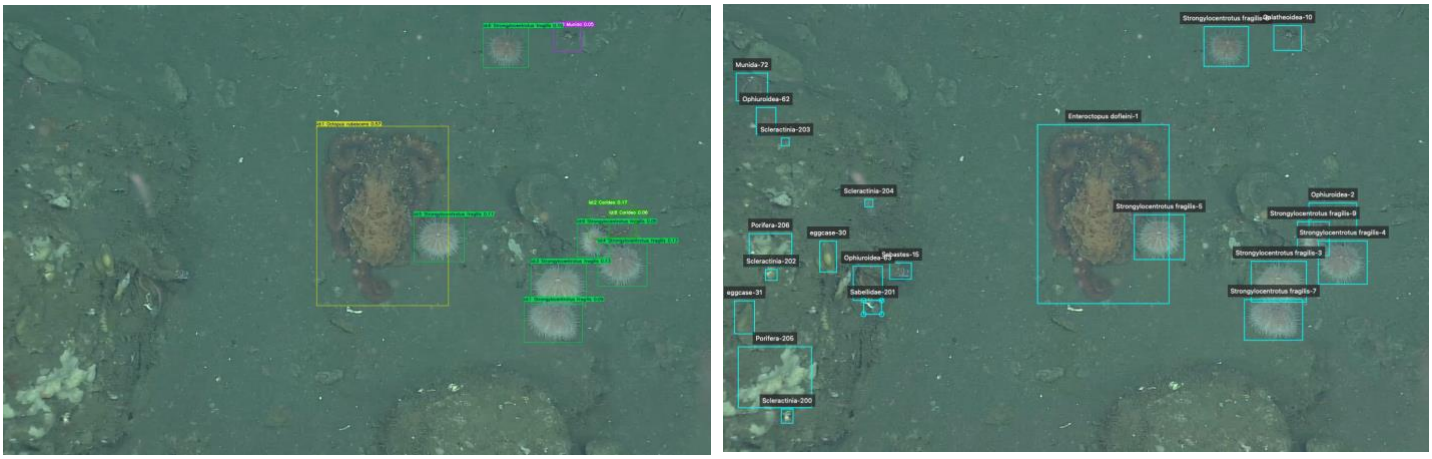
In this study, we aim to model and extend the work of AnimalTrack to multi-animal tracking in the deep-sea environment. Working in this setting provides further challenges, as there is often low visibility, near constant camera and/or vehicle motion, and high occlusion. This benchmark fills a critical gap in MBARI's Video Lab annotation pipeline, which previously lacked a standardized metric for evaluating model tracker performance. With this benchmark, model and tracker results can be quantified and compared against internal detector and tracker changes and, also, with results reported in the literature.

3. Materials and Methods

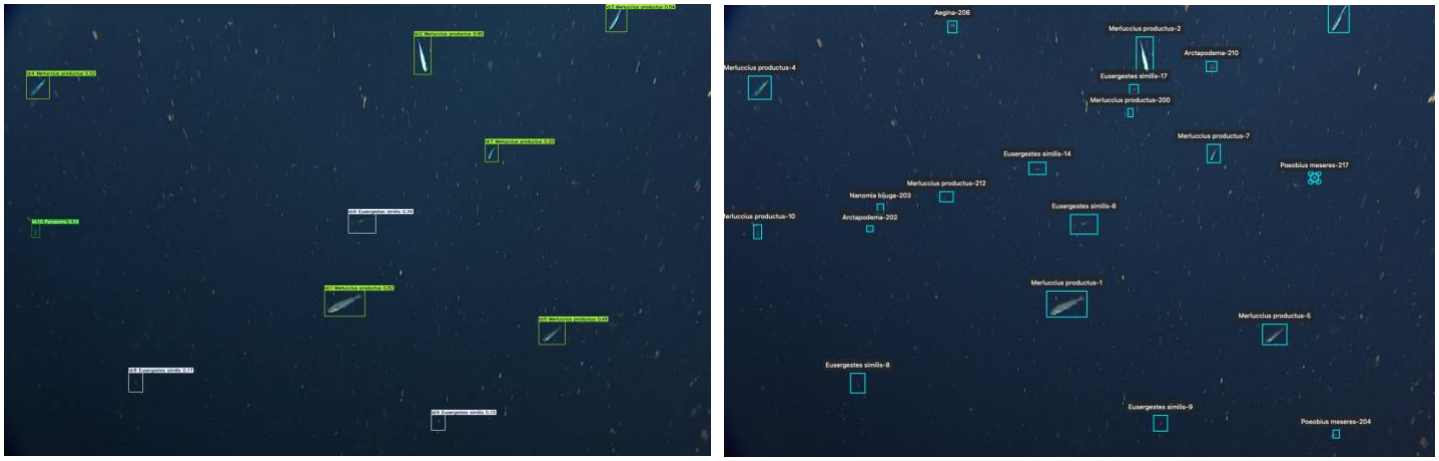
Model Prediction | Ground Truth



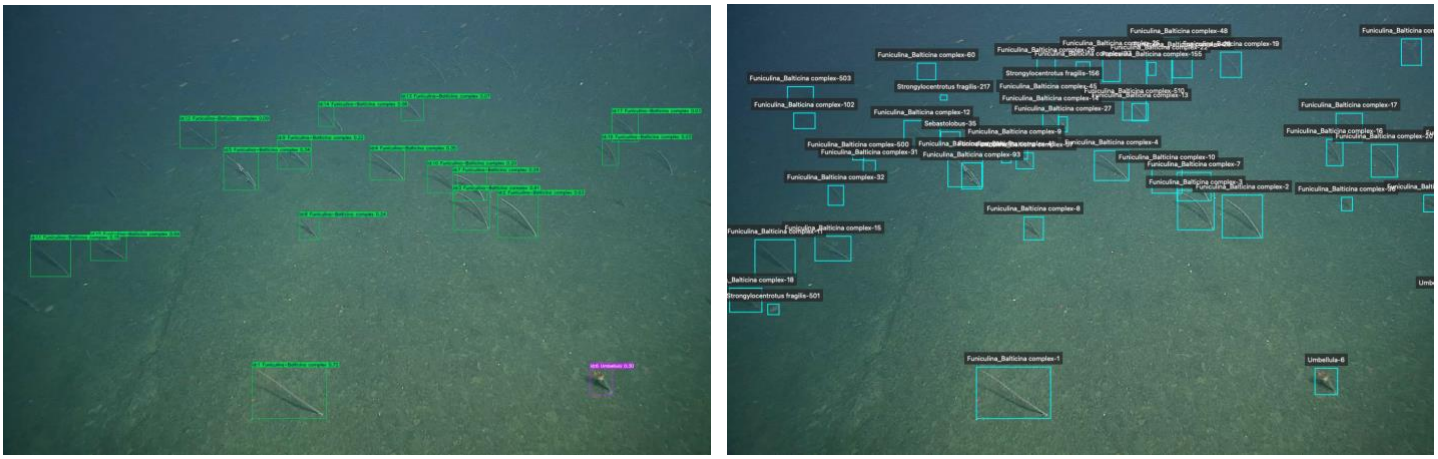
Simple Midwater



Simple Benthic



Difficult Midwater



Difficult Benthic

Fig 1. Model prediction and ground truth detection screenshots (first frame) for each video sequence

A gold standard benchmark dataset was developed to evaluate model and tracker performance. As shown in Fig. 1, it consists of four video sequences: two videos for the midwater and two videos for the benthic environment. Each environment includes a simple and difficult video sequence, with difficulty defined largely by the number of animals present. All videos are 10 seconds long and have a resolution of 60 frames per second. To generate the ground truth files, the video was first run through the model and tracker to produce preliminary bounding boxes. These results were then imported to RectLabel, an image annotation tool, to manually fit localizations to the object or create them if it wasn't detected by the model. After all 600 frames were reviewed for each video sequence, the resulting ground truth files were directly compared to the model

prediction files for evaluation. Fig. 2 illustrates the number of ground truth detections and unique IDs in each video sequence.

Vid Sequence	# of Detections	# of Unique IDs
Simple Midwater	1945	8
Simple Benthic	13864	29
Difficult Midwater	12940	57
Difficult Benthic	28658	94

Fig 2. Number of ground truth detections and IDs in each video sequence

The object detection model used in the study is a YOLOv8 model trained on 452k training localizations of 415 class labels. The model parameters were previously optimized for accuracy with MBARI's typical imagery from this unique environment. The main parameters adjusted were confidence threshold, which was set to 0.025, and the IoU threshold, which was set to 0.4. The tracker used in this study is the ByteTrack (Zhang et al., 2022) tracker implemented by Ultralytics ("Multi-Object Tracking with Ultralytics YOLO", 2025). The tracker was also previously fine-tuned using extensive visual analysis and iterative testing of various combinations of hyperparameter configurations by Kris Walz.

The evaluation metric chosen was the HOTA metric, which is an accuracy score that equally weights detection, localization, and association. These three components are all calculated using an IoU (intersection over union) score. For instance, localization IoU is calculated as the overlap between the ground truth and predicted detection boxes over the union of the two boxes. The total localization accuracy is an average of all localization IoU scores. A similar formula is used to compute detection and association accuracy, and all three values are combined into a single, unified HOTA score (Luiten et al., 2021). HOTA is made up of multiple submetrics, such as accuracy, precision, and recall for both detection and association, as well as localization accuracy. This decomposition enables a clearer understanding of the strengths and limitations of the model and tracker.

The *TrackEval: HOTA (and other) evaluation metrics for Multi-Object Tracking (MOT)* Github repository was used to compute the HOTA metric for the benchmark videos (Luiten et al., 2021). This repository was used as a framework for the codebase in the overall benchmarking workflow. All code was written in Python and implemented in a Python 3.9.23 environment using Ultralytics YOLO version 8.3.166 and PyTorch version 2.2.2 on a CPU (Apple M4 architecture).

4. Results

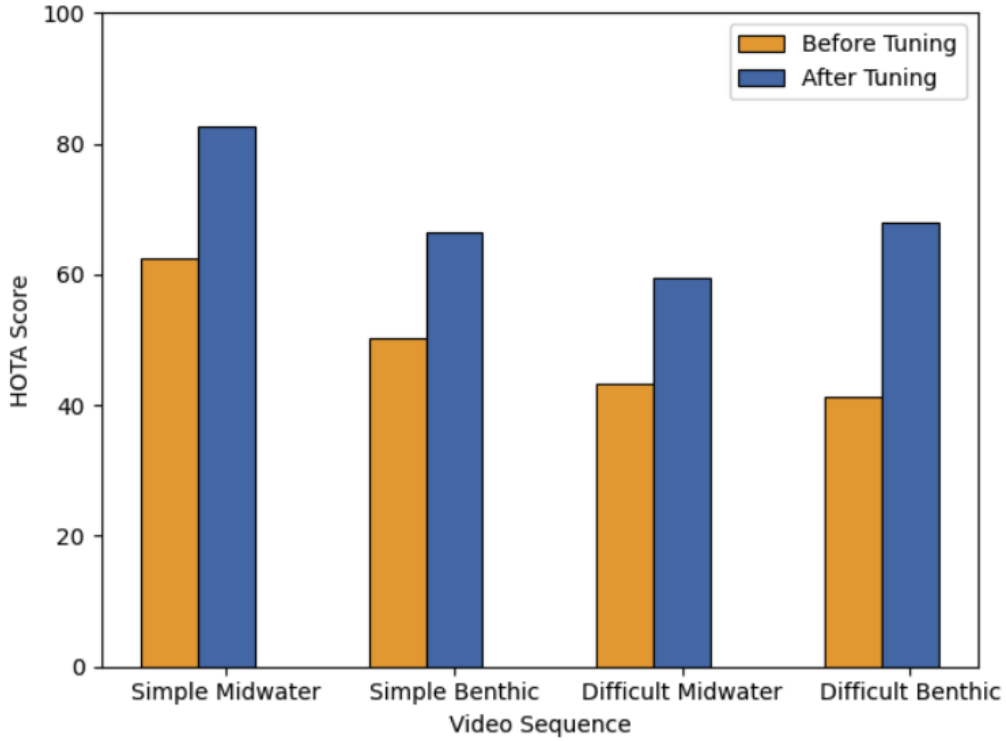


Fig 3. Bar chart comparing HOTA scores before and after tracker parameter tuning

The figure above presents the HOTA scores for the 4 evaluated video sequences. The model and tracker achieved the best performance on the simple midwater video, with an accuracy of 82.69%. This outcome was consistent with our initial assumptions, as there were only a few animals to detect and were highly visible for most of the video. Difficult benthic ranked second with a score of 68.06% and simple benthic followed that with an accuracy of 66.44%. The lowest performance was observed with the difficult

benthic video at 59.49% accuracy. This was also anticipated due to high occlusion in the background and the quick movement of the camera down the midwater column. In particular, animals such as *Poeobius meseres* were often missed, as they appeared in the video as subtle shadows in the background.

Fig. 3 also compares the HOTA scores of the tracker before and after hyperparameter tuning. The untuned tracker uses the default settings from Ultralytics ("Multi-Object Tracking with Ultralytics YOLO", 2025) and showed much lower performance on average - approximately 20% less compared to the tuned tracker. As mentioned earlier, the tracker was tuned manually by careful visual review and iterative hyperparameter adjustments. In this case, metrics were used to report the results but were not used to tune the hyperparameters. The significant improvement in HOTA scores highlights the value of using domain knowledge and expertise in optimizing model and tracker performance.

Following model and tracker evaluation, an attempt was made to further optimize the tracker using information gathered from the HOTA metric. The current tracker, ByteTrack, was compared against Bot-SORT, a tracker also supported by Ultralytics. Hyperparameter configurations were tested on the difficult midwater video sequence and parameters that were adjusted include the thresholds for first and second-stage associations, new track initialization, matching tracks, etc. Despite using supplemental information gathered from HOTA sub-metrics, such as recall and precision for detection and association, no improvements in accuracy were achieved, suggesting that the tracker was already highly optimized. This is supported by the already significant differences in accuracy between the default and tuned tracker results seen in Fig. 3.

In addition to the codebase, a comprehensive documentation was created to support future development, such as adding videos to the benchmark dataset or implementing new trackers to improve performance. It can be accessed publicly through the MBARI GitHub organization and follows the documentation format for many of their existing tools. It includes instructions for installation, usage, and helpful tips such as an explanation of directory file structure and file examples.

5. Discussion

Based on these results, we conclude that footage from the benthic versus midwater environment does not have a significant impact on performance, as had been assumed by Video Lab staff. Instead, the content of the video, such as number of animals, camera motion, and occlusion have a greater impact on detection and tracking accuracy. This is supported by the fact that difficult benthic performed better than simple benthic, as seen in the results. In the simple benthic video, the camera initially remains stationary, then pans to the side before returning to the original location. This movement caused the tracker to assign new IDs to the same animals when they temporarily left the frame, which largely reduced the HOTA score. On the other hand, the difficult benthic video had more animals present, but were predominantly sea pens, which had minimal movement and were thus easier to track. These factors combined led to difficult benthic achieving higher accuracy, underscoring the significance of video content on model performance.

Future work could include implementing new trackers, in addition to ByteTrack (Zhang et al., 2022) and Bot-SORT (Aharon et al., 2022), to attempt to improve model performance. We tested the BoxMOT repository (Broström, 2023), which included a range of high performing trackers such as BoostTrack (Stanojevic et al., 2024), StrongSORT (Du et al., 2023), OC-SORT (Cao et al., 2023), Deep OC-SORT (Maggiolino et al., 2023), ByteTrack, and Bot-SORT. However, after evaluation, HOTA scores were significantly lower, even for ByteTrack and Bot-SORT. Reasons for this discrepancy are unclear but could be due to differences in internal tracker implementations. Next steps could also explore expanding the benchmark dataset to incorporate video sequences with greater diversity and complexity. This would enable a more well-rounded evaluation of model performance and help identify the conditions where the model performs better or worse. Once available, this work could be used for a Kaggle type competition as well.

6. Conclusions

In this study, we developed a benchmark for MBARI's Video Lab annotation pipeline. The benchmark dataset featured simple and difficult video sequences of the

midwater and benthic environments. We used the HOTA metric for evaluation of detection, localization, and association accuracy. A key takeaway is that while quantitative metrics, such as HOTA, can be a useful for evaluating object detection models and tracker performance, they must be balanced by including domain expertise. Specifically, extensive visual analysis is essential for interpreting results and validating model performance in real world applications. Thus, more exchange between domain experts and data scientists is critical for developing robust methods for evaluating model and tracker performance.

7. ACKNOWLEDGEMENTS

I'd like to thank Lonny Lundsten and everyone in the Video Lab - Kris Walz, Kyra Schlining, Nancy Jacobsen-Stout, Larissa Lemon, Giovanna Sainz, Megan Bassett, Kvein Barnard, and Brian Schlining - for their help and support this summer. I'd like to acknowledge Kris Walz for helping with midwater annotations and tracker hyperparameter tuning, and Kevin Barnard for helping with documentation and code reorganization. Special thanks to George Matsumoto and Megan Bassett for organizing this internship and creating such a unique opportunity. The MBARI Summer Internship Program is generously supported through a gift from the Dean and Helen Witter Family Fund and the Rentschler Family Fund in memory of former MBARI board member Frank Roberts (1920-2019) and by the David and Lucile Packard Foundation. Additional funding is provided by the Maxwell/Hanrahan Foundation.

References:

Aharon, Nir and Orfaig, Roy and Bobrovsky, Ben-Zion. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv preprint arXiv*:2206.14651 (2022). https://doi.org/10.48550/arXiv.2206.14651

Broström, Mikel. BoxMOT: pluggable SOTA tracking modules for object detection, segmentation and pose estimation models. *Zenodo* (2023). https://zenodo.org/record/7629840

Cao, Jinkun and Pang, Jiangmiao and Weng, Xinshuo and Khirodka, Rawal and Kitani, Kris. Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023).
https://doi.org/10.48550/arXiv.2203.14360

Du, Yunhao and Zhao, Zhicheng and Song, Yang and Zhao, Yanyun and Su, Fei and Gong, Tao and Meng, Hongying. StrongSORT: Make DeepSORT Great Again. *IEEE Transactions on Multimedia*, vol. 25, pp. 8725-8737 (2023).
https://doi.org/10.48550/arXiv.2202.13514

Luiten, J., Ošep, A., Dendorfer, P. *et al.* HOTA: A Higher Order Metric for Evaluating Multi-object Tracking. *Int J Comput Vis* 129, 548–578 (2021).
https://doi.org/10.1007/s11263-020-01375-2

Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. AnimalTrack: A Benchmark for Multi-Animal Tracking in the Wild. *Int J Comput Vis* 131, 496–513 (2023).
https://doi.org/10.48550/arXiv.2205.00158

Maggiolino, Gerard and Ahmad, Adnan and Cao, Jinkun and Kitani, Kris. Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification. *arXic preprint arXiv:*2302.11813 (2023).
https://doi.org/10.48550/arXiv.2302.11813

"Multi-Object Tracking with Ultralytics YOLO." *Ultralytics YOLO Docs*, 22 June 2025, docs.ultralytics.com/modes/track/. Accessed 12 July 2025.

Schlining, B.M., Jacobsen Stout, N. 2006. MBARI's video annotation and reference system. In: Proceedings of the Marine Technology Society/Institute of Electrical and Electronics Engineers Oceans Conference, Boston, MA, pp. 1–5.

Stanojevic, Vukasin D and Todorovi, Branimir T. BoostTrack: boosting the similarity measure and detection confidence for improved multiple object tracking. *Machine Vision and Applications*, 0932-8092 (2024).
https://doi.org/10.48550/arXiv.2408.13003

Zhang, Yifu and Sun, Peize and Jiang, Yi and Yu, Dongdong and Weng, Fucheng and Yuan, Zehuan and Luo, Ping and Liu, Wenyu and Wang, Xinggang. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. *Proceedings of the European Conference on Computer Vision* (2022).
https://arxiv.org/abs/2110.06864