People Tracking Methods Applied to Planalto Palace Security Videos

Cristiano B. de Oliveira*, João C. Neves†, Rafael O. Ribeiro‡ and David Menotti§
*Federal University of Ceará, Quixadá, Brazil
†University of Beira Interior, NOVA-LINCS, Portugal
‡National Institute of Criminalistics, Brazilian Federal Police, Brasília, Brazil
*§Department of Informatics, Federal University of Paraná, Curitiba, Brazil
* cristianobac@ufc.br † jcneves@di.ubi.pt ‡ rafael.ror@pf.gov.br § menotti@inf.ufpr.br

***Abstract*—This paper presents a work in progress with comparative results for five state-of-the-art approaches for pedestrian tracking (Deep OC-SORT, OC-SORT, StrongSORT, BotSORT and ByteTrack) applied to a preliminary version of the UFPR-Planalto801 dataset, composed by footage taken from security cameras in Palácio do Planalto, the official office of the President of Brazil. The videos show images of the protesters invasion occurred on January 8, 2023. We used pieces of the public released footage in order to conduct the experiments in a real-world context. The trackers were evaluated by using IDF1, CLEAR and HOTA metrics. The results show a large number of ID switches and missed associations, and a maximum HOTA score of 0.46, achieved by StrongSORT and ByteTrack methods, which shows how challenging is this type of scenario.**

I. Introduction

The ubiquitous presence of cameras noticed in recent years, whether the ones installed for monitoring and surveillance purposes or those integrated into personal devices, has led to an unprecedented scenario of visual data capturing human activities and interactions. This surge in visual data has prompted the need for advanced techniques to analyze, understand, and extract meaningful information from videos, which may be used for several purposes, such as person identification and pedestrian tracking.

In this sense, video-based pedestrian tracking, which refers to the process of automatically detecting and following individuals in video, is specially useful for applications such as autonomous driving, urban planning, surveillance and crowd management. Furthermore, it may also assist in investigations by providing visual evidence of incidents and capturing the movements of potential suspects. By detecting and tracking individuals in crowded spaces, security systems can identify suspicious behavior, monitor unauthorized access, and, consequently, contribute to enhance public safety.

As widely reported by the media, the Palácio do Planalto, the official office of the President of Brazil, was infiltrated by protesters on January 8, 2023. The videos[1] of the surveillance cameras inside the building were made public available by brazilian's Supreme Court and released on April 23, 2023 [1]. The images are of high quality, and the videos collectively comprise over 1TB of data. Due to the nature of the images, which depict a real-world surveillance scenario, there has arisen an interest in utilizing such footage to construct a dataset for use in the development of security systems.

Thus, this paper presents comparative results for a number of approaches regarding to pedestrian tracking applied to a preliminary version of a dataset composed by part of these security videos. Section III-A presents more details about the dataset. To the best of our knowledge there is no other similar comparison conducted involving this particular footage, despite a similar work regarding action detection in surveillance scenarios being presented in [2].

For the purpose of understand the characteristics and challenges of this scenario, we selected five state-of-the-art trackers for evaluation: **Deep OC-SORT** [3], **OC-SORT** [4], **StrongSORT** [5], **BotSORT** [6] and **ByteTrack** [7]. These particular trackers have performed well when applied to datasets like DanceTrack [8] and MOTChallenge [9], [10], commonly used in Multi-Object Tracking (MOT) research.

We evaluated the results by using IDF1 [11], CLEAR [12] and HOTA [13] metrics. The trackers scored pretty similar values for the adopted metrics, but, overall, ByteTrack [7] achieved slightly better results. The results also show a large number of ID switches and missed associations for all trackers.

II. Tracking by Detection

Tracking by Detection (TBD) is a common paradigm to address the challenges posed by Multiple Object Tracking (MOT) tasks. TBD approaches for people tracking typically consist in two major steps: 1) detecting individuals in each frame of a video sequence; and 2) associating these detections with unique identifiers. These steps centrally involve the linkage of detected people across consecutive frames to ultimately establish coherent trajectories.

As an intermediary step, TBD models usually include some technique for modeling the temporal coherence of object movements in order to improve the prediction and correction of object trajectories. Techniques based on Kalman filters are common choices to motion prediction.

As occurs in many Computer Vision related applications, problems like occlusions, poor light conditions, abrupt motion changes, etc., help to decrease the efficiency of the models.

[1]Online available in https://drive.presidencia.gov.br/public/615ba7

This results in missing detections and missing/wrong associations, which are key problems that still persist in MOT. These problems can lead algorithms to perform identity switches and to compute wrong trajectories. Crowded scenes are prone to this, as are scenarios where an individual leaves the scene and subsequently reappears after a period of time.

A. Trackers

ByteTrack [7] is a tracking by association method that uses every bounding box detected in order to determine the objects tracklets, on the premise that low score detection boxes are due mainly to occlusions. It performs two association steps. In the first step the algorithm considers only high scores boxes and uses Intersection over Union (IOU) or a Re-ID feature distance between the detected and predicted bounding boxes. For the second step it uses the low score detection boxes and computes their similarities with the unmatched tracklets in order to filter out the background detections and recover true objects to be matched.

BoT-SORT [6] is a tracking algorithm derived from SORT [14]. As occurs in most SORT-like algorithm, Bot-SORT adopts the Kalman filter as a motion model to predict tracklets through frames. However, Bot-SORT applies slightly changes in the state vector, by inserting values for width and height in order to improve the fit of the bounding boxes. It also includes a camera motion compensation module based on optical flow and uses IOU in combination with appearance Re-ID descriptors, obtained by a ResNet50 backbone network. As well as in ByteTrack, Bot-SORT uses two steps for associating detected bounding boxes to tracklets.

StrongSORT [5] is another SORT-like algorithm and it is built upon DeepSORT [15], one of the first methods to use deep learning techniques for tracking. There are a few differences regarding DeepSORT. At first, it relies on YOLOX [16] as an strong detector, therefore the name StrongSORT. It also includes an EMA (Exponential Moving Average) module to improve long-term association and changes the motion model to a NSA Kalman Filter based [17]. Camera movement is compensated by a enhanced correlation coefficient maximization technique.

OC-SORT [4] aims to improve robustness during occlusion and non-linear motion, when targets have non-constant velocities within a time interval. For such purpose, the authors proposed an Observation-Centric Re-Update (ORU) strategy, which generates a re-updated version of the Kalman Filter based on historical observation for periods when targets are not tracked. The authors claims that the use of these observations reduce the accumulated error over time. Thus, objects that are not tracked for a period of time are revisited and tracklets can be re-activated. To tackle the no-linear motion problem, OC-SORT considers the direction of motion when performing associations. This is done by adding term named Observation-Centric Momentum (OCM) to the association cost matrix.

Deep OC-SORT [3] evolved from OC-SORT by dynamically inserting visual appearance into the tracking model and then improving the accuracy and robustness of associations. Deep OC-SORT uses low detector confidence to identify situations like occlusion or blur, and rejects them when computing the similarity cost. Therefore, the process is adapted in such way that it increases the weight of appearance features only in cases of high-quality detections.

III. Experiments

A. Data preparation

For this paper we built a dataset, named UFPR-Planalto801, composed with indoor footage taken inside the Palácio do Planalto, during rioters invasion on January 8, 2023. The whole set of released videos comprises over 1TB of data. The videos recorded the full day and several locations inside the palace. They are organized into 33 folders, with 1557 videos in total, with a resolution of 1920x1080 pixels, mostly recorded at 100 frames per second (FPS).

The invasion occurred on a Sunday and, therefore, rooms were typically unoccupied and most images taken before the invasion occurs shows nothing but empty rooms. Thus, due to the huge amount of available data, we have selected and clipped parts of videos showing some people activity after the invasion time (around 3 pm). These clipped videos were then re-encoded to 24 FPS, in order to reduce the number of frames to be processed.

A fair part of the sampled footage shows chaotic/complex scenes, as shown in Fig. 1. Theses scenes include people in different rooms and angles. Several individuals are using similar clothes and/or accessories like masks, coats, hats, bags, flags, or even holding wood bars (Fig. 1a and 1b) . Other situations include the presence of smoke (Fig. 1c) and people images reflected on glasses, as in Fig. 1d.

People detection for clipping the videos was done by using YOLO (You Only Look Once) [18], a well-known models family for detecting objects in images. Since its first release YOLO evolved into different versions [19] [20]. In this paper we refer to YOLOv8 [21], which also supports tracking. In order to annotate the frames we have used as references the bounding boxes of pedestrians detected by YOLOv8, which were manually corrected as needed. Source code for trackers are provided by [22]. We used pre-trained models in all experiments setups.

B. Evaluation Metrics

For evaluation we used the Higher Order Tracking Accuracy (HOTA) metrics [13], the CLEAR Metrics Multi-Object Tracking Accuracy (MOTA) and Multi-Object Tracking Precision (MOTP) [12], and the ID F1 Score (IDF1) [11]. These metrics were calculated by using the TrackEval [23] codebase, in such way that the higher the value, the better the tracker performs.

MOTA calculates the relationship between the total ground-truth detections and the number of false positives, false negatives and ID switches, while MOTP calculates the total average error in relation to the estimated and detected position. IDF1 aims to measure the quality of predicted trajectories, based on the association of IDs over the trajectories. HOTA is a general metric that comprises a set of sub-metrics derived from MOTA

(a)

(b)

(c)

(d)

Fig. 1. Examples of scenes captured in footage. (a) and (b) Images from rooms 1 and 2, respectively; (c) Smoke in room 3; (d) Glass reflections in room 3.

and MOTP. The main sub-metrics are detection accuracy (DetA) and association accuracy (AssA), which comprise their respective values for recall and precision. These sub-metrics are useful to differentiate between errors related to detection or association.

IV. Preliminary Results

Table I presents a summary of the scores (HOTA, CLEAR and IDF1) for each tracker, while Table II presents the number of detections and IDs related to the ground-truth (GT). The numbers of ID switches and trajectories fragments is presented in Table III.

TABLE I
Summary of Results for HOTA, CLEAR and IDF1

	BotSORT	ByteTrack	Deep OC-SORT	OC-SORT	StrongSORT
HOTA	0.451	**0.467**	0.450	0.450	0.464
MOTA	0.555	0.600	0.608	0.610	**0.612**
MOTP	0.815	**0.826**	0.815	0.815	0.815
IDF1	0.505	**0.524**	0.496	0.495	0.521

TABLE II
Number of Detections and IDs

Tracker	# of Detections	# of IDs
BotSORT	112665	814
ByteTrack	105229	730
Deep OC-SORT	105525	790
OC-SORT	105554	721
StrongSORT	104968	959
Ground-Truth	122881	272

ByteTrack and StrongSORT are the trackers with the best HOTA and IDF1 scores in comparison to the other trackers. Respectively, the former achieved **0.467** and **0.524** while the latter achieved 0.464 for HOTA and 0.521 for IDF1. The results for ByteTrack also show it is the second tracker with the largest number of missed detections, missing 14.36% from the 122881 detections in GT. Despite of that, it achieved the best MOTP score, which is **0.826**. All the remaining trackers achieved a MOTP of 0.815. MOTA values are 0.600 for ByteTrack and **0.612** for StrongSORT. ByteTrack is also the tracker with the smallest number of ID switches, with 475 in total.

The computed HOTA score for Deep OC-SORT (0.450) was similar to those achieved by OC-SORT (also 0.450) and BotSORT (0.451). However, the number of ID switches when using Deep OC-SORT is the highest, performing 799 ID switches, as one can see in Table III. BotSORT predicted 814 IDs, while there are only 272 in GT, and performed 514 ID switches, resulting in an IDF1 score of 0.505.

MOTA and IDF1 for Deep OC-SORT are 0.608 and 0.496, respectively. Both these results are very close to the ones for OC-SORT, which achieved 0.610 and 0.495, for MOTA and IDF1, respectively. Despite of this, OC-SORT shows a number of IDs (721) closer to GT and also a lower number of ID switches (613) in comparison to Deep OC-SORT.

By observing the detections we noticed that tracks were frequently fragmented into short ones, with the same individual being associated to several IDs along the video. Thus, we

TABLE III
NUMBER OF ID SWITCHES AND FRAGMENTS

Tracker	# of ID Switches	# of Fragments
BotSORT	514	1793
ByteTrack	475	1749
Deep OC-SORT	799	1939
OC-SORT	613	1929
StrongSORT	589	1855

believe that a potential improvement in this particular context would be the inclusion of a subsequent step to associate tracklets fragments.

Another interesting result is the high number of associated IDs by each tracker in comparison to the number of GT IDs (please see Table II). This shows that all tested approaches have failed to maintain a consistent ID over time.

There are mainly two cases for ID switching. First case is when two or more individuals are associated to the same ID. Figure 2a shows a situation that can confuse the motion prediction and the appearance modules of the trackers. In such situations, one person leaves the scene while another one enters and is detected in a similar pose and in approximate coordinates, misleading the tracker to associate the same ID. A second case occurs when an individual departs from the scene and reappears moments later. In such cases the predominant outcome is the assignment of a new ID to the individual, consequently initiating a new tracklet, like shown in figures 2b and 2c.

Figures 3 to 7 show the values for HOTA sub-metrics. As one can see in these figures, the major differences among the trackers are related to the association step, with an AssA of 0.38 for ByteTrack, and 0.35 for both Deep OC-SORT and OC-SORT.

V. CONCLUSION

This paper presented preliminary comparative results for a number of approaches regarding people tracking. For such purpose we created a dataset, from a real-world scenario, containing public security videos with indoor footage taken inside the Palácio do Planalto, during protesting acts occurred on Jan. 8, 2023.

Since the videos in dataset comprehend long duration footage, we have mainly selected approaches that claim to improve motion models over time. However, as results shows, tracking motion in long-length videos still remains problematic.

As a work in progress, this paper does not intend to diminish the significance or criticize the employed methods negatively; rather, it aims to utilize them as a foundation for constructing a proposal that addresses the issues present in the utilized dataset. Furthermore, we intend to expand the dataset with new images and conduct additional tests in the future.

(a)

(b)

(c)

Fig. 2. Examples of wrongly associated IDs. (a) Two persons wrongly associated to the same ID; (b) and (c) Same person associated to different IDs.

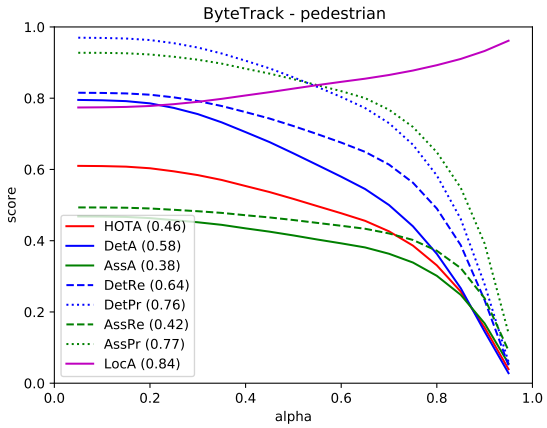

Fig. 3. HOTA Results for ByteTrack

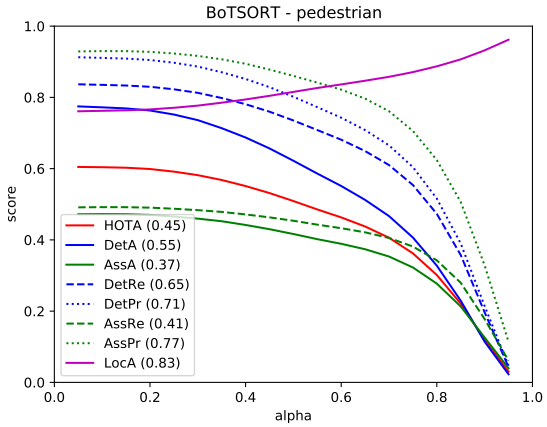

Fig. 4. HOTA Results for BotSORT

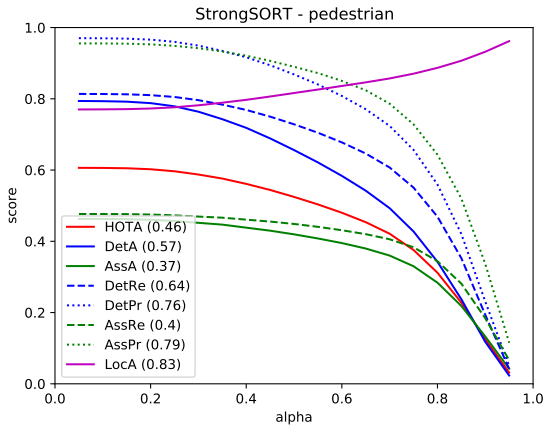

Fig. 5. HOTA Results for StrongSORT

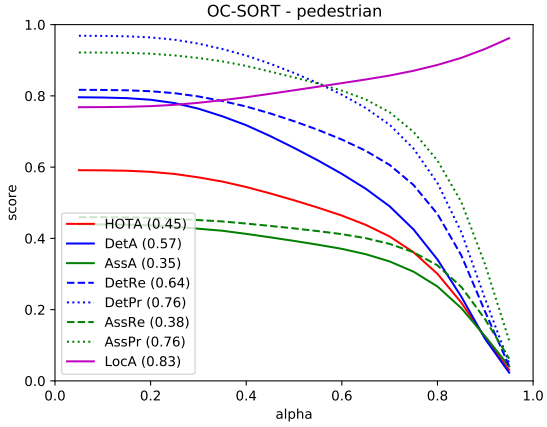

Fig. 6. HOTA Results for OC-SORT

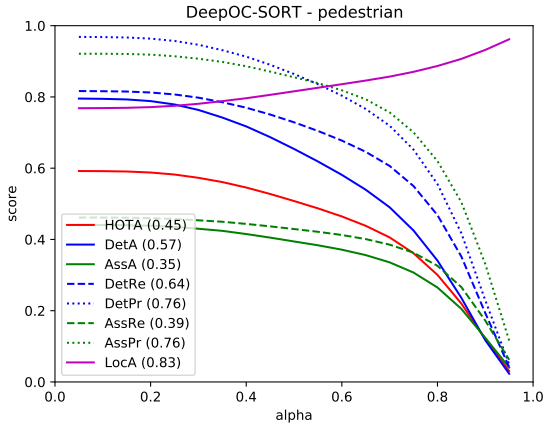

Fig. 7. HOTA Results for Deep OC-SORT

A. Future Work

To enhance the outcomes, it is our intention to assess models employing approaches that incorporate biometric information. Many frames encompass facial features, which would enable accurate individual re-identification through facial recognition. As an alternative, we can apply re-identification via gait recognition, particularly in cases where facial visibility is limited.

Given the substantial volume of frames within this specific dataset, the task of annotating the images demands considerable effort. Hence, it is also our interest to employ unsupervised learning techniques to facilitate the utilization of a larger portion of the footage.

ACKNOWLEDGMENTS

This work was supported in part by the Coordination for the Improvement of Higher Education Personnel (CAPES) (*Programa de Cooperação Acadêmica em Segurança Pública e Ciências Forenses # 88881.516265/2020-01*), and in part by the National Council for Scientific and Technological Development(CNPq) (# 308879/2020-1). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro RTX 8000 GPU used for this research.

REFERENCES

[1] Gabinete de Segurança Institucional, "Nota à imprensa," 2023, last accessed 25 September 2023. [Online]. Available: https://www.gov.br/gsi/pt-br/centrais-de-conteudo/noticias/2023-1/nota-a-imprensa-acesso-as-imagens-do-dia-08-01-2023-do-circuito-interno-de-seguranca-do-palacio-do-planalto

[2] J. Alikhanov and H. Kim, "Online Action Detection in Surveillance Scenarios: A Comprehensive Review and Comparative Study of State-of-the-Art Multi-Object Tracking Methods," *IEEE Access*, vol. 11, pp. 68 079–68 092, 2023.

[3] G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani, "Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification," *arXiv preprint arXiv:2302.11813*, 2023.

[4] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. 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)*, June 2023, pp. 9686–9696.

[5] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "StrongSORT: Make DeepSORT Great Again," *IEEE Transactions on Multimedia*, 2023.

[6] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[7] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[8] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, and P. Luo, "DanceTrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 20 993–21 002.

[9] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "CVPR19 tracking and detection challenge: How crowded can it get?" *arXiv:1906.04567 [cs]*, Jun. 2019, arXiv: 1906.04567. [Online]. Available: http://arxiv.org/abs/1906.04567

[10] ——, "MOT20: A benchmark for multi object tracking in crowded scenes," *arXiv:2003.09003[cs]*, Mar. 2020, arXiv: 2003.09003. [Online]. Available: http://arxiv.org/abs/1906.04567

[11] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 17–35.

[12] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *J. Image Video Process.*, vol. 2008, jan 2008. [Online]. Available: https://doi.org/10.1155/2008/246309

[13] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking," *International Journal of Computer Vision*, pp. 1–31, 2020.

[14] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464–3468.

[15] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3645–3649.

[16] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," 2021.

[17] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, and J. Dong, "GIAOTracker: A comprehensive framework for MCMOT with global information and optimizing strategies in VisDrone 2021," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 2809–2819.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo Algorithm Developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922001363

[20] J. Terven and D. Cordova-Esparza, "A Comprehensive Review of YOLO: From YOLOv1 and Beyond," 2023.

[21] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[22] M. Broström, "BoxMOT: A collection of SOTA real-time, multi-object trackers for object detectors ," 2023. [Online]. Available: https://github.com/mikel-brostrom/yolo_tracking

[23] A. H. Jonathon Luiten, "TrackEval," https://github.com/JonathonLuiten/TrackEval, 2020.