






Comparative Analysis of Deep Learning-Based Multi-Object Tracking Approaches Applied to Sports User-Generated Videos

Elton Alencar¹^a, Larissa Pessoa¹^b, Fernanda Costa²^c, Guilherme Souza²^d and Rosiane de Freitas²^e

¹Programa de Pós-Graduação em Informática (PPGI), UFAM, Manaus-Amazonas, Brazil

²Universidade Federal do Amazonas (UFAM), Manaus-Amazonas, Brazil

Keywords: DeepSORT, Deep Learning, Mobile Devices, StrongSORT, TrackFormer, YOLO-World, YouTube, Zero-Shot Tracker.

Abstract: The growth of video-sharing platforms has led to a significant increase in audiovisual content production, especially from mobile devices like smartphones. Sports user-generated videos (UGVs) pose unique challenges for automated analysis due to variations in image quality, diverse camera angles, and fast-moving objects. This paper presents a comparative qualitative analysis of multiple object tracking (MOT) techniques applied to sports UGVs. We evaluated three approaches: DeepSORT, StrongSORT, and TrackFormer, representing detection and attention-based tracking paradigms. Additionally, we propose integrating StrongSORT with YOLO-World, an open-vocabulary detector, to improve tracking by reducing irrelevant object detection and focusing on key elements such as players and balls. To assess the techniques, we developed UVY, a custom sports UGV database, having YouTube as its data source. A qualitative analysis of the results from applying the different tracking methods to UVY-Track videos revealed that tracking-by-detection techniques, DeepSORT and StrongSORT, performed better at tracking relevant sports objects than TrackFormer, which focus on pedestrians. The new StrongSORT version with YOLO-World showed promise by detecting fewer irrelevant objects. These findings suggest that integrating open-vocabulary detectors into MOT models can significantly improve sports UGV analysis. This work contributes to developing more effective and scalable solutions for object tracking in sports videos.


1 INTRODUCTION


Video content, a multi-modal structure, has gained significant importance as an efficient means of sharing information, often surpassing traditional media composed of text and images. The rapid expansion of video-sharing platforms (e.g., social media) in recent years has contributed to an exponential increase in video production, with millions of videos being generated daily (Tang et al., 2023). In addition, most of these platforms have mobile devices, such as smartphones, as their main source of data generation and consumption. This is primarily because smartphones have built-in cameras that enable quick video capture


(Wang et al., 2023).


Most of these videos, recorded using handheld cameras on mobile devices, are **User-Generated Videos (UGVs)** (Guggenberger, 2023). When multiple UGVs capture the same event, a multi-perspective view is created, such as in a football stadium. However, manually processing such large volumes of videos remains a time-consuming and labor-intensive task, creating a growing demand for automated tools for analysis and management. To meet this demand, deep learning-based video understanding methods and analysis technologies have emerged, leveraging intelligent analysis techniques to automatically recognize, extract and interpret video features, significantly reducing manual workload (Tang et al., 2023).


In this context, **Multi-Object Tracking (MOT)** is one of the key tasks in **video understanding**, as it allows for the continuous monitoring of entities within a video frame (Wang et al., 2024). Currently, deep learning techniques are widely applied in object

^a  <https://orcid.org/0000-0002-2610-7071>

^b  <https://orcid.org/0000-0002-8307-6443>

^c  <https://orcid.org/0009-0000-6702-7222>

^d  <https://orcid.org/0009-0000-1113-9348>

^e  <https://orcid.org/0000-0002-7608-2052>

detection and tracking systems (Russell and Norvig, 2020). This capability contributes to the automation of video content understanding and a variation of tracking features-based analysis (Amosa et al., 2023), which facilitates the application in different knowledge areas, such as sports video analysis (Rangasamy et al., 2020), action recognition (Alencar et al., 2022), video summarization, video synchronization (Whitehead et al., 2005). These applications demonstrate the versatility and growing importance of MOT in diverse fields of knowledge (Amosa et al., 2023).

There are various MOT approaches, which can be categorized into four paradigms: tracking-by-detection, tracking-by-regression, tracking-by-segmentation, and tracking-by-attention (Meinhardt et al., 2022). Among these paradigms, **Tracking-by-Detection** (TBD) has become the most explored paradigm due to the advances in deep learning-based object detection approaches (Amosa et al., 2023), (Du et al., 2023). Similar to what is illustrated in Figure 1, the process in this paradigm begins by detecting the objects of interest. A unique identifier is then assigned to each detected object, and its location is propagated across subsequent frames using a model that maintains object associations throughout the video (Ishikawa et al., 2021).

One of the most well-established tracking-by-detection models in the literature is **StrongSORT** (Du et al., 2023), an improved version of DeepSORT (Wojke et al., 2017), which was built on top of SORT (Simple Online and Realtime Tracking) (Bewley et al., 2016), a classic MOT method that predicts an object's current position based on its previous location. The motion prediction in these processes is achieved by matching detection bounding-boxes with predicted positions, relying on the NSA Kalman Filter and Hungarian matching for optimization (Du et al., 2023).

Additionally, as shown in Table 1, YOLO (and its more than 8 versions) has been used for detecting objects to be tracked (Hussain, 2024). Most of these versions, including the latest one, have publicly available pre-trained weights that were trained using the MS COCO dataset, which contains a limited number of object categories (Lin et al., 2014). This limitation restricts their ability to detect objects in sports videos, where elements often fall outside the predefined categories.

The introduction of solutions such as **YOLO-World (2024)**, which features open-vocabulary detection capabilities, mitigates this limitation by improving the YOLOv8 architecture by integrating a text-encoder based vision-language models (Cheng et al., 2024). This enables the detection of objects not pre-

viously categorized. In summary, as illustrated in the Figure 1 (II), it comprises three main components: (1) YOLOv8, used as the detector model to extract multi-scale features from input images; (2) a CLIP-based text encoder that converts text into embeddings; and (3) a custom network that performs multi-level cross-modality fusion between image features and text embeddings.

2 PROBLEM DEFINITION

Most recent tracking algorithms primarily focus on pedestrian or vehicle tracking (Ishikawa et al., 2021)(Huang et al., 2024). These algorithms have shown significant progress in public benchmarks like MOT16, MOTS20, and MOT20 (Dendorfer et al., 2020). However, they face significant challenges in sports scenarios (Huang et al., 2024). Sports videos are characterized by fast movements, frequent occlusions, and constant changes in perspective, which requires specific solutions for automatic analysis, such as tactical analysis and performance evaluation of athletes. The inability of MOT algorithms to adapt to these challenges highlights the need for more robust approaches (Zhao et al., 2023).

In addition to these challenges, the use of models like YOLO requires effort for fine-tuning or retraining when adapting to specific sports objects, such as players or balls. While fine-tuned models can achieve high performance for a defined set of classes, this process demands considerable resources, particularly when dealing with UGVs that feature diverse and unpredictable conditions. Open-vocabulary object detectors, such as YOLO-World, offer an alternative by reducing the need for re-training, as they are designed to generalize across a broader set of object categories (Cheng et al., 2024).

Given these limitations, this paper presents a qualitative and comparative analysis of multi-object tracking techniques. Specifically, we evaluate tracking-by-detection methods (*i.e.*, DeepSORT, StrongSORT) and tracking-by-attention methods (*i.e.*, TrackFormer) on user-generated videos recorded in sports events. The qualitative analysis between these approaches will allow us not only to identify the effectiveness of each technique in relation to the limitations described, but also to propose advances in the field of object tracking in challenging sports scenarios.

Figure 1 illustrates the complete workflow of the main processes for the proposed analysis. The input frames extracted from UGV videos are processed through the YOLO-World object detection module.

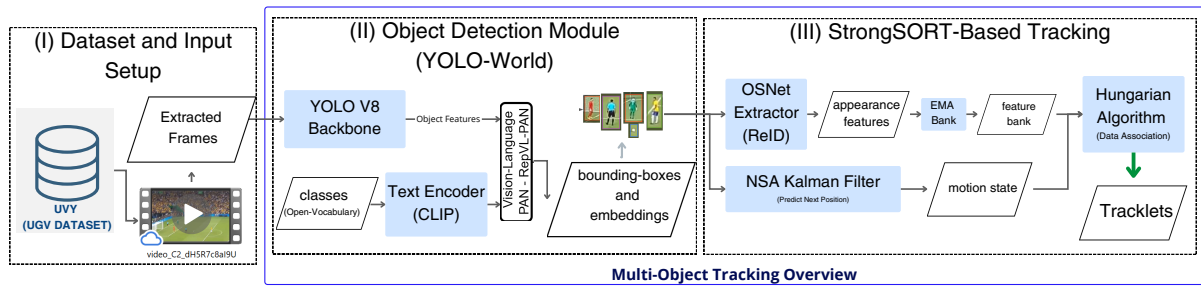


Figure 1: Overview of the proposed multi-object tracking framework for UGVs recorded in sport events.

The detected bounding-boxes and embeddings are passed into the StrongSORT-based tracking, where appearance features (extracted via OSNet) and motion states (predicted by the NSA Kalman Filter) are used to perform robust data association using the Hungarian Algorithm. This integration aims to enable a more robust and adaptable approach to object tracking in sports scenarios, particularly when dealing with UGVs.

2.1 Related Works

The Table 1 presents a chronological organization of key works on multi-object tracking. These works demonstrate the evolution of deep learning-based MOT algorithms, which initially focused on controlled environments, such as pedestrian tracking with static cameras (Ishikawa et al., 2021). Over time, these methods have been extended to more complex scenarios, such as observing animals application (Dolokov et al., 2023), sport scenarios (Huang et al., 2024), including multi-view camera setups (Cherdchusakulchai et al., 2024). All the studies listed in the table below emphasize that tracking is a fundamental task in the field of computer vision (Huang et al., 2023).

Most of the works listed follow the tracking-by-detection approach. Initially, the listed deep-learning-based algorithms were applied and evaluated in controlled environments, where static cameras recorded videos for tasks like pedestrian tracking (e.g., MOT17, MOT20, MOTS20 dataset) (Ishikawa et al., 2021). This is evident in works such as “TrackFormer”, “MOTRv2”, and “StrongSORT”. The former introduced innovations with a tracking-by-attention approach, which uses a Transformer-based model with attention mechanisms and adaptive filtering to improve object association between video frames.

As tracking demands increased, particularly for dynamic environments like sports and user-generated videos, new challenges emerged. According to (Huang et al., 2023), the process of object tracking

in sports scenarios presents two main challenges: (1) the nonlinear movement of players and (2) the similar appearance of athletes on the field. Thus, tracking objects in more unpredictable environments can present unique challenges. In the work proposed by (Huang et al., 2024), these challenges are approached by replacing the Kalman filter with an iterative ExpansionIoU technique and deep features. Despite these advances, methods like this can still face significant limitations when applied to UGVs, where variable capture conditions, such as lighting and camera angles, add further challenges to tracking.

Furthermore, it was observed that, of all the studies listed in Table 1, only one emphasizes the importance and challenges posed by UGVs. In professional sports broadcasts, high-quality cameras are used to record videos in high resolution and with a high frame rate, combined with image processing for referee assistance or data collection. However, this requires more resources. Therefore, developing a solution with low resource requirements for data collection could be significant, given the abundance of videos in this context (Huang et al., 2019). For these reasons, during the development of this work, the qualitative analysis was prioritized for the performance of MOT models based on detection and attention applied to user-generated sports videos.

3 EXPERIMENTAL PROTOCOL

This section outlines the steps used to conduct the comparative study between tracking-by-detection and tracking-by-attention techniques. It includes the characteristics and assumptions considered for the creation of the dataset, the methods implemented, and, finally, the stages followed for the comparative analysis. The research is characterized as an experimental and descriptive study with a comparative design. The main objective is to compare the performance of publicly, pre-trained, state-of-the-art MOT algorithms when applied to the task of object tracking in user-generated sports videos.

Table 1: List of works related to the multi-object tracking, highlighting their evaluation datasets and approaches.

Multi-Object Tracking Works (None UGV-based)		
Title & Reference	Evaluation Dataset	Approaches
Analysis of Recent Re-Id Architectures for Tracking-by-Detection Paradigm in MOT (Ishikawa et al., 2021)	MOT20 (pedestrians).	TBD approach. Comparative analysis of the quantitative results for DeepSORT when replacing the Re-ID process.
TrackFormer: Multi-Object Tracking with Transformers (Meinhardt et al., 2022)	MOT17 and MOTS20 (pedestrians).	Tracking-by-attention: Transformer-based. Feature extraction: ResNet-50 (CNN).
StrongSORT: Make DeepSORT Great Again (Du et al., 2023)	MOT17 and MOT20.	TBD paradigm. Detector: YOLOX-X. ReID Feature extraction: BoT + EMA. Prediction: NSA Kalman Filter.
Upper Bound Tracker: A Multi-Animal Tracking Solution for Closed Laboratory Settings (Dolokov et al., 2023)	MultiTracker Mice Custom Annotation.	TBD paradigm, using OC-SORT as baseline. Detector: YOLOX-X.
Online Multi-camera People Tracking with Spatial-temporal Mechanism and Anchor-feature Hierarchical Clustering (Cherdchusakulchai et al., 2024)	2024 AI City Challenge Track 1 (synthetic scenes).	MOT: YOLOv8 + OSNet (Re-ID) + ByteTrack. MTMC: Merges tracklets.
GMT: A Robust Global Association Model for Multi-Target Multi-Camera Tracking (Fan et al., 2024)	VisionTrack (recorded by drones).	Tracking-by-attention: Global MTMC. Detector: CenterNet. Feature extraction: Re-ID (Mask R-CNN). Association: Hungarian algorithm.
Iterative Scale-Up ExpansionIoU and Deep Features Association for Multi-Object Tracking in Sports (Huang et al., 2024)	SportsMOT and SoccerNet-Tracking (players).	Tracking-by-detection (sports scenarios). Detector: YOLOX + OSNet (Re-ID). Track prediction: Expansion IoU.

3.1 UVY Dataset

No existing dataset met the requirements of the study, specifically user-generated sports match recordings captured by mobile devices. Most public benchmarks, such as MOT16, MOTS20, and MOT20, focus on pedestrian or vehicle tracking. Therefore, a custom dataset—the UVY-Track—was created to evaluate MOT models on user-generated videos. The UVY dataset (User-generated Videos from YouTube) was developed following a structured pipeline that mirrors the steps used in the creation of the MUVY dataset (Pessoa et al., 2024), ensuring a format inspired by benchmarks like MOT16 and MOT20 (Dendorfer et al., 2020).

Fifteen user-generated sports videos—four basketball, six volleyball, and five soccer—were selected from YouTube, prioritizing varying quality, mobile device recordings, and durations under four minutes, made publicly available via Creative Commons licenses. A Python script automated the download and metadata extraction processes, retrieving information such as video ID, title, URL, and duration using public libraries like OpenCV. Frames were extracted using FFmpeg and organized into folders named by source video and frame position.

To streamline annotation, the process transitioned from manual labeling using Google’s Vertex AI to a hybrid approach combining automatic detection with YOLO-World and manual validation. The detection model was used specifically to assist in obtaining bounding boxes of objects it can detect, leveraging its zero-shot learning capability to identify objects across diverse and unstructured contexts available in UGVs.

At this stage, the dataset includes only the bounding box regions of detected objects, without assigning a unique identifier for each object throughout the video. A manual validation process complemented the automatic detection to ensure the quality of the annotations, including correcting or refining the detected bounding boxes and annotating objects missed by YOLO-World. Each video is stored in a uniquely named folder containing original and YOLO-processed frames, detected object metadata, and .mp4 files. The dataset is publicly available on Zenodo¹.

3.2 Algorithm Implementation

This phase focused on researching and selecting state-of-the-art deep learning-based MOT techniques for implementation and reproduction. Based on the analysis in Section 2.1, priority was given to models capable of tracking multiple objects simultaneously. Pre-trained models were chosen, as the goal was to validate its capability of tracking sports objects in UGVs, rather than training new MOT models. Adaptations were made to ensure that the selected models could handle sports-specific scenarios, tracking only relevant objects, such as players and sports balls, while maintaining object identity across frames. The chosen algorithms—DeepSORT, StrongSORT, and TrackFormer—were prioritized for their ease of implementation and strong community support.

The detection models (YOLOv5 and YOLOv7)

¹UVY-Track

were applied to each frame of the videos to generate bounding boxes around detected objects. The tracking models (DeepSORT, StrongSORT, and TrackFormer) were then used to assign unique identifiers and maintain object trajectories across frames. Both detection and tracking models were iteratively tested with different thresholds and parameters, such as confidence scores and association metrics, adjusted to improve tracking accuracy. DeepSORT leverages Kalman filtering and the Hungarian algorithm for robust tracking in complex scenarios (Wojke et al., 2017). StrongSORT enhances DeepSORT by improving occlusion handling and track consistency (Du et al., 2023). Finally, TrackFormer, an attention-based method, explores alternative paradigms for object tracking (Meinhardt et al., 2022).

The implementation followed instructions from each tool's GitHub repository. To avoid issues with versioning dependencies of the Python libraries used by each algorithm, the entire execution process was carried out in the online environment Google Colaboratory through the creation of notebooks for each tested algorithm/tool². This also ensured the correct use of the PyTorch library and its dependencies, which required GPUs.

3.3 Tracking-by-Detection: StrongSORT

Initially, to understand the implementation process of object tracking techniques, the steps for DeepSORT were followed using the Kaggle-Code-Repository publication (Pareek, 2022). After setting up access to the video inputs, the detector model was configured to obtain the tuple containing (bounding box location[left, top, w, h], confidence, detectedClass), extracted from each object present in each frame. The tuple was input into the DeepSORT model, which performed estimation, association, and Tracker ID lifecycle tasks. The same process was applied to implement StrongSORT, a MOT algorithm that integrates three techniques for improved performance.

StrongSORT uses YOLOv7 (Wang et al., 2022) for accurate and fast object detection. Its robust data association combines appearance and motion information to maintain detections across frames, even under occlusions and appearance variations. Additionally, a deep neural network enables object re-identification when they disappear and reappear in the scene (Du et al., 2023). The implementation followed the instructions from the GitHub repository (Du et al., 2023), and a Colab notebook was created. The process was validated by successfully processing the sports videos from the dataset using both Deep-

SORT and StrongSORT. These results can be seen in Videos 01 and Video 03 available at the link below².

3.4 Tracking-by-Attention: TrackFormer

TrackFormer is a MOT algorithm that uses the concept of Transformer models. Its main contribution lies in the application of self-attention to learn long-range representations and make precise data associations between consecutive frames. This approach makes it possible to capture contextual dependencies between different objects in the scene, resulting in tracking that is more robust to occlusions, changes in appearance and the entry of new objects (Meinhardt et al., 2022). The implementation followed the instructions provided in the authors' GitHub repository (Meinhardt et al., 2022). A Google Colab notebook was created for execution, and the implementation was validated by processing the sports videos from the created dataset, including Video 02, which can be accessed at the link below².

4 RESULTS AND ANALYSIS

From the process of selecting and implementing multiple object tracking techniques, three models were implemented (DeepSORT-2017, TrackFormer-2022, and StrongSORT-2023) and applied to videos from the dataset described in the previous section (Figure 2). Initially, the analysis of these preliminary results was entirely qualitative, but in the future we intend to evaluate the results by a quantitative analysis of the models when applied to the UVY dataset using usual MOT performance metrics (e.g. MOTA, HOTA, etc).

During the processing of the videos selected from the new database, and from a **comparative and qualitative analysis** of the results observed in the output videos returned from each approach, it was possible to observe that factors such as the quality of the video, the speed of the objects, and the complexity of the scene can influence the performance of the algorithms, which was already expected, since the user-generated videos have varied recording conditions.

Furthermore, Figure 2 illustrates that DeepSORT and StrongSORT successfully detected the ball. This was due to the fact that YOLOv5 and YOLOv7 were pre-trained models specialized to detect objects such as "sports ball". In contrast, TrackFormer focused on detecting pedestrians or people, as its CNN-based architecture was designed for extracting human-based

²Output videos folder.

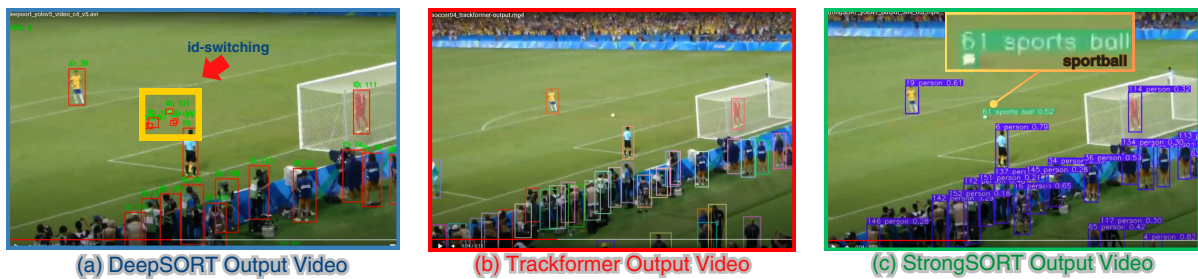


Figure 2: Comparison of MOT techniques: (a) DeepSORT with ID-switch issues; (b) TrackFormer missing “sports ball”; (c) StrongSORT detecting “sports ball” after improving data association.

features. (Meinhardt et al., 2022). This observation highlighted the possibility of adapting tracking-by-detection methods (e.g., DeepSORT and StrongSORT) by replacing the detector by a model capable of detecting objects of interest according to the proposed context.

Now, comparing the results between DeepSORT (YOLOv5) and StrongSORT (YOLOv7), it was confirmed that the former suffers from the id-switch problem during the data association process, *i.e.*, for the same detected object, such as the ball, DeepSORT assigns more than one id (tracklet-ID). However, this had already been mentioned by the authors of StrongSORT (an improved version of DeepSORT) (Du et al., 2023).

4.1 NEW: StrongSORT with YOLO-World (2024)

The implemented StrongSORT (YOLOv7) presented promising results that indicated its effectiveness in tracking objects in UGV sports videos. However, thinking in the context of a sports analysis application, where you only want to observe the trajectory of the player and actor in action on the field (Zhao et al., 2023), it was observed the necessity of a model that only tracks objects that are common between the different videos, which means that for sports videos, the ideal is to carry out the analysis process only for objects that are on the field (*i.e.*, “person playing”, “sports ball”, “referee”, “goalkeeper”)

For this reason, a new version of StrongSORT was implemented, replacing the detector model. Previously, YOLO-v7 was used. The updated version now tracks objects detected by YOLO-World. Figure 3 presents a side-by-side results: the top image shows a video processed with the existing StrongSORT+YOLOv7, while the bottom shows the same video processed by the new StrongSORT+YOLO-World, developed in this research. Without requiring model training, this version effectively reduces

the number of tracked objects, focusing mainly on the active objects in the field.

Despite this satisfactory result for the context of this work, it’s important to mention that this new version will still have cases in which it detects an object that is outside the field of action, such as the case highlighted (yellow) in the bottom image of Figure 3. However, when compared in terms of numbers for this specific case, the number of objects detected by the new model is around four times less than the number detected by the original model. This can be confirmed in video 14, available at the link below³.

Therefore, based on the qualitative analysis presented above, it was concluded that the new adapted deep learning-based multi-object tracker, StrongSORT+YOLO-World, can be used to track relevant objects detected in sports videos recorded by users in order to obtain enough visual features, extracted automatically, for a sports analysis, based on the user’s perspective.

Additionally, in order to confirm how this new version of StrongSORT (with YOLO-World), introduced in this work, can be considered as an **open-vocabulary multi-object tracker**, since it enables the detection and tracking of objects not previously categorized, experiments were reproduced for random video UGVs outside the sports context, in order to confirm whether this new tracker can successfully track new object classes. An example of a successful result for this can be seen in the Video 13 (available at the link below³, where the model was asked to track objects classified as [“snake”, “feet”, “basket”, “flute”]. For the majority of the time, it was able to detect the new object class and assign a unique id to track it over time. Multiple unique object IDs were assigned for consistent tracking throughout the video, confirming its ability to detect and track object classes outside predefined categories.

³Output videos folder.



Figure 3: Side-by-side comparison of StrongSORT with YOLOv7 (a) and StrongSORT with YOLO-World (b). The YOLO-World integration enables custom class detection (e.g. 'person playing', 'audience') without retraining in sports scenarios.

5 CONCLUDING REMARKS

This work presented a comparative analysis of tracking approaches (DeepSORT, StrongSORT, and TrackFormer), revealing that tracking-by-detection models performed better in user-generated sports videos than the tracking-by-attention model, TrackFormer. For the context presented, where pre-trained state-of-the-art models were evaluated without retraining, TrackFormer showed inferior performance, particularly in detecting the ball. However, retraining such models for specific scenarios could potentially improve their performance.

Additionally, the introduction of a novel approach, StrongSORT integrated with YOLO-World (an open-vocabulary detector), improved the tracking capabilities by focusing on relevant objects and reducing noise from irrelevant objects. This demonstrates the utility of open-vocabulary models in reducing effort for training detection model. However, specialized models trained to detect specific classes in sports UGVs could achieve similar or even superior results.

The dataset introduced, UYV-Track, is in its initial version and has limitations regarding the number of videos and manual effort required for labeling. Future work aims to address these limitations by automating parts of the dataset population process, including the use of Large Language Models (LLMs) to assist in the classification of user-generated sports videos.

In conclusion, the study confirms that deep learning-based MOT methods, particularly those with detection models adapted to the sports scenario, can improve tracking performance in UGVs. These findings contribute to the development of robust tools for automated sports analysis, including a UGV dataset, paving the way for future work to quantitatively evaluate these methods, explore further adaptations, and expand the applicability of these approaches to broader and more complex scenarios.

ACKNOWLEDGMENTS

This work is part of the PD&I SWPERFI Project (AI Techniques for Software Performance Analysis, Testing, and Optimization), a partnership between UFAM and MOTOROLA MOBILITY, with members from the ALGOX research group (Algorithms, Optimization, and Computational Complexity) of CNPq (National Council for Scientific and Technological Development - Brazil). It also receives support by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Finance Code 001, and is partially supported by Amazonas State Research Support Foundation - FAPEAM - through the POSGRAD project 2024/2025.

REFERENCES

- Alencar, E. D. N. d. A. et al. (2022). Extração de características de narrativas audiovisuais a partir de elementos visuais e suas relações.
- Amosa, T. I., Sebastian, P., Izhar, L. I., Ibrahim, O., Ayinla, L. S., Bahashwan, A. A., Bala, A., and Samaila, Y. A. (2023). Multi-camera multi-object tracking: a review of current trends and future advances. *Neurocomputing*, 552:126558.
- Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE.
- Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., and Shan, Y. (2024). Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911.
- Cherdchusakulchai, R., Phimsiri, S., Trairattanapa, V., Tungjitnob, S., Kudisthalert, W., Kiawjak, P., Thamwiwatthana, E., Borisuitsawat, P., Tosawadi, T., Choppradi, P., et al. (2024). Online multi-camera people tracking with spatial-temporal mechanism and anchor-feature hierarchical clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7198–7207.

- Dendorfer, P., Rezatofghi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., and Leal-Taixé, L. (2020). Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.
- Dolokov, A., Andresen, N., Hohlbaum, K., Thöne-Reineke, C., Lewejohann, L., and Hellwich, O. (2023). Upper bound tracker: A multi-animal tracking solution for closed laboratory settings. In *VISIGRAPP (5: VISAPP)*, pages 945–952.
- Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., and Meng, H. (2023). Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*.
- Fan, H., Zhao, T., Wang, Q., Fan, B., Tang, Y., and Liu, L. (2024). Gmt: A robust global association model for multi-target multi-camera tracking. *arXiv preprint arXiv:2407.01007*.
- Guggenberger, M. (2023). *Multimodal Alignment of Videos*. Doctoral dissertation, Alpen-Adria-Universität Klagenfurt, Klagenfurt am Wörthersee. Toward Multimodal Synchronization of User-Generated Event Recordings.
- Huang, H.-W., Yang, C.-Y., Ramkumar, S., Huang, C.-I., Hwang, J.-N., Kim, P.-K., Lee, K., and Kim, K. (2023). Observation centric and central distance recovery for athlete tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 454–460.
- Huang, H.-W., Yang, C.-Y., Sun, J., Kim, P.-K., Kim, K.-J., Lee, K., Huang, C.-I., and Hwang, J.-N. (2024). Iterative scale-up expansion and deep features association for multi-object tracking in sports. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 163–172.
- Huang, Y.-C., Liao, I.-N., Chen, C.-H., İk, T.-U., and Peng, W.-C. (2019). Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE.
- Hussain, M. (2024). Yolov1 to v8: Unveiling each variant—a comprehensive review of yolo. *IEEE Access*, 12:42816–42833.
- Ishikawa, H., Hayashi, M., Phan, T. H., Yamamoto, K., Masuda, M., and Aoki, Y. (2021). Analysis of recent re-identification architectures for tracking-by-detection paradigm in multi-object tracking. In *VISIGRAPP (5: VISAPP)*, pages 234–244.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Meinhardt, T., Kirillov, A., Leal-Taixe, L., and Feichtenhofer, C. (2022). Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854.
- Pareek, N. (2022). Using deepsort object tracker with yolov5. Kaggle. [Online]. Available: <https://www.kaggle.com/code/nityampareek/using-deepsort-object-tracker-with-yolov5>. Accessed: Dec. 10, 2024.
- Pessoa, L., Alencar, E., Costa, F., Souza, G., and Freitas, R. (2024). Exploring multi-camera views from user-generated sports videos. In *Anais do XII Symposium on Knowledge Discovery, Mining and Learning*, pages 105–112, Porto Alegre, RS, Brasil. SBC.
- Rangasamy, K., As’ari, M. A., Rahmad, N. A., Ghazali, N. F., and Ismail, S. (2020). Deep learning in sport video analysis: a review. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 18(4):1926–1933.
- Russell, S. and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. 4th edition.
- Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., et al. (2023). Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.
- Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.
- Wang, J., Chen, D., Luo, C., He, B., Yuan, L., Wu, Z., and Jiang, Y.-G. (2024). Omnivid: A generative framework for universal video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18209–18220.
- Wang, Y., Huang, Q., Jiang, C., Liu, J., Shang, M., and Miao, Z. (2023). Video stabilization: A comprehensive survey. *Neurocomputing*, 516:205–230.
- Whitehead, A., Laganiere, R., and Bose, P. (2005). Temporal synchronization of video sequences in theory and in practice. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION’05)-Volume 1*, volume 2, pages 132–137. IEEE.
- Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE.
- Zhao, Z., Chai, W., Hao, S., Hu, W., Wang, G., Cao, S., Song, M., Hwang, J.-N., and Wang, G. (2023). A survey of deep learning in sports applications: Perception, comprehension, and decision. *arXiv preprint arXiv:2307.03353*.