Exploring multi-camera views from user-generated sports videos

Larissa Pessoa[1], Elton Alencar[1], Fernanda Costa[2], Guilherme Souza[2], Rosiane de Freitas[2]

[1] Programa de Pós-Graduação em Informática (PPGI/UFAM)
[2] Universidade Federal do Amazonas (UFAM)
lsp,enda,fernanda.costa,guilherme.souza,rosiane@icomp.ufam.edu.br

Abstract.

The proliferation of mobile devices with video recording capabilities has revolutionized audiovisual content creation, sharing, and consumption, turning user-generated video (UGV) platforms into major data sources. Despite this growth, there is a notable gap in publicly available datasets featuring multiangle recordings of sports events captured with various mobile cameras. This paper introduces the MUVY Dataset, which offers a diverse collection of sports videos from multiple perspectives, unrestricted by video size. The dataset addresses common challenges in user-generated videos, such as shaking, occlusions, blurring, and abrupt movements. Each video is accompanied by metadata that include camera identification, YouTube URLs, extracted frames, and object annotations. Covering sports like soccer, American football, artistic gymnastics, athletics, basketball, tennis, and cricket, the MUVY Dataset facilitates advancements in video understanding and viewpoint selection. Initial experiments in camera pose estimation demonstrate the dataset's potential for training models in this domain. Additionally, it supports the selection of the closest viewpoint based on object detection and the relative area occupied by detected objects. Overall, the MUVY Dataset aims to advance multi-camera video analysis and related research areas.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: multicam, object detection, sport events, video dataset, user-generated video, Youtube.

1. INTRODUCTION

The widespread adoption of mobile devices with advanced video recording capabilities has driven the creation and consumption of video content. Platforms like YouTube and TikTok have become popular social networks, attracting users with their video content [Zhang et al. 2023]. This increase in video sharing has transformed these platforms into valuable data repositories, advancing machine learning model training in video analysis. This paper introduces the MUVY dataset, comprising recordings from multiple cameras, offering various angles and perspectives crucial for applications like object tracking and optimal view selection.

This simulates human visual perception using videos recorded on mobile devices, allowing for complex analyses in sports scenarios. Publishing the dataset aims to advance video analysis techniques for mobile-recorded and multi-camera videos. The structure of this article is as follows. Section 2 overviews theoretical foundations involving video understanding, multiple camera views, user-generated videos, camera pose estimation, object detection, and object tracking. Section 3 defines the problem and motivation for creating the dataset and reviews relevant studies. Section 4 details the methodology used to construct the dataset. Section 5 describes the dataset's characteristics. Section 6 discusses three possible applications of the dataset to evaluate its suitability. Finally, the article concludes with a summary and future steps.

2. THEORETICAL BACKGROUND

This section explores the fundamental concepts underpinning this research, including video understanding and applied concepts necessary for dataset validation experiments.

Video understanding combines computer vision, machine learning, and AI to automate video analysis, enabling systems to identify objects, recognize actions, and interpret events. This process reflects human perception in interpreting complex visual narratives and is closely related to the field of Computer Vision, which uses image processing, machine learning, and pattern recognition techniques to simulate human vision [Tang et al. 2023].

User-generated videos (UGV), created by individuals rather than media companies, are popular on platforms like YouTube, TikTok, and Twitch. Handheld devices, especially smartphones and tablets, are essential for UGV creation and distribution, democratizing video production with advanced cameras and accessible editing software [Naab and Sehl 2017]. However, these videos often face issues such as shakiness and occlusions, challenges that deep learning techniques seek to mitigate [Su et al. 2017].

Multiple camera views systems use multiple cameras to capture the same scene from different angles, providing a more comprehensive understanding and aiding in 3D reconstruction. This method is widely applied in areas such as surveillance, sports broadcasting, and autonomous navigation, despite challenges related to synchronization and data fusion [Olagoke et al. 2020].

Camera pose estimation is crucial in computer vision and photogrammetry for applications like augmented reality, robotic navigation, 3D reconstruction, and environmental mapping. It determines a camera's position and orientation in 3D space by aligning images with a global coordinate framework, ensuring accurate overlay of virtual elements and precise navigation. Deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized pose estimation, providing robust alternatives to traditional methods by effectively learning correlations between scene inputs and camera poses, even in complex environments [Citraro et al. 2020].

Object detection and tracking models detect and track objects frame by frame in videos. Advances in object detection and the prevalence of the Tracking-by-Detection paradigm in multi-object tracking (MOT) techniques have been well documented [Cheng et al. 2024]. MOT is widely used for automating video content understanding, including applications in sports video analysis, action recognition, and summarization. Progress has been significant on benchmarks like MOT16, MOTS, and MOT20, though challenges persist with sports-specific datasets. Object tracking in sports videos aids in automatic analysis, tactical evaluation, and athlete movement statistics, while feature-based tracking supports multiple video analyses [Zhao et al. 2023].

3. RELATED WORKS

The growth in multimedia content generation and sharing has made the creation of datasets with extracted features a common practice in computer vision and machine learning. The diversity of scenarios in YouTube videos has enabled the creation of such datasets [Cho et al. 2023][Deliege et al. 2021]. Consequently, video datasets have become crucial in scientific research, focusing on various aspects and data types. Several multi-camera view datasets have been developed for various purposes. Table I provides an overview of seven notable datasets, highlighting their limitations compared to the proposed one.

The work [Perera et al. 2020] provides 2324 videos, of which only 150 are sports-related clips, specifically American football, with multiple perspectives obtained from YouTube, but these clips are limited in length and rely exclusively on official broadcasts. Similarly, the MultiSports v1.0 dataset [Li et al. 2021] collects sports videos from YouTube but does not guarantee multi-camera views.

The [Cricri et al. 2013] dataset is a nonpublic UGV dataset with multi-camera videos from sporting events recorded with mobile devices. User-generated videos is present in the Multi-sensor Concert Recording Dataset [Bailer et al. 2015] and the Jiku Mobile Video Dataset [Saini et al. 2013], but they lack real-world challenges because the participants know they are being recorded for the experiment.

Table I. Comparison of Datasets on Sports or with Multi-Camera Views.

Dataset	Reference	Videos	Duration (min)	Avg Video Length	Multi-Views	Source	Domain	Application
Multi-Viewpoint Outdoor Action Recognition	[Perera et al. 2020]	2324	280	7 sec	yes	YouTube & Drone	Humans	Action Recognition
UGV in Public Sport Events	[Cricri et al. 2013]	507	4382	8 min	yes	Mobile Cameras	Multi Sports	Sport Genre Classification
MultiSports v1.0	[Li et al. 2021]	3200	1115	20 sec	no	YouTube	Multi Sports	Action detection
Multi-sensor Concert Recording	[Bailer et al. 2015]	160	-	-	yes	Static and Mobile Cameras	Concerts	Concert Classification
Jiku Mobile Video	[Saini et al. 2013]	473	1841	4 min	yes	Mobile Cameras	Concerts	Video mashups
DukeMTMC	[Ristani et al. 2016]	8	680	85 min	yes	Static Cameras	Pedestrian	MTMCTracking [Wang et al. 2019]
SoccerNet-MVFouls	[Held et al. 2023]	8923	744	5 sec	yes	Official Broadcasts [Giancola et al. 2018]	Soccer	Foul Classification
MUVY	(ours)	141	195	83 sec	yes	YouTube	Multi Sports	Cam Pose Estimation Viewpoint Selection Multi-View Tracking

In contrast, the MUVY Dataset, by extracting videos from YouTube, captures spontaneous sports events from the fans' perspective, offering a rich variety of content for analysis. Other datasets offer multi-camera views but have limitations: DukeMTMC [Ristani et al. 2016] focuses on a single outdoor scene with fixed cameras, limiting its applicability to broader contexts; SoccerNet-MVFouls [Held et al. 2023] is restricted to short clips from official broadcasts [Giancola et al. 2018] [Deliege et al. 2021].

As previously mentioned, there are multi-view datasets with videos recorded in the same environment, but none use YouTube for user-generated videos recorded simultaneously at sports matches. Additionally, the number of videos alone is not sufficient. For example, SoccerNet-MVFouls has many videos, but their average duration of 5 seconds is unsuitable for analyses requiring a greater temporal context. Recent studies highlight the versatility and impact of data mining techniques. These methods optimize model compression for YOLOv3 [de Aguiar Salvi and Barros 2020], balancing efficiency and performance for processing large video datasets [Gonçalves et al. 2019]. These techniques can be adapted for annotating and classifying multi-view sports videos in our dataset.

4. DATASET CONSTRUCTION METHODOLOGY

A multi-camera views dataset, MUVY *(available for download* [1]*)*, was constructed by adapting principles from the Knowledge Discovery in Databases (KDD) process [Fayyad 1997], as illustrated in Figure 1. The following steps were systematically followed to achieve this:

(1) **Selection:** Initially, the search was manually conducted to find Creative Commons-licensed videos on YouTube, focusing on those recorded simultaneously during the same sporting event. Queries included sport genre, specific actions (e.g., goal, penalty), and competition season (e.g., World Cup, Olympics). Priority was given to user-generated videos recorded from different viewpoints. The selected videos were manually organized into YouTube playlists to ensure multi-view features. Queries were adjusted to find events with at least two videos capturing the same action simultaneously. Each playlist represents a specific sporting event with different perspectives.
(2) **Preprocessing**: The playlists are organized in a CSV file to allow manual adjustment of the URLs, ensuring they can be correctly read by the extraction script. Data cleaning removed duplicates, corrected inconsistencies, and handled missing information. A Python script automated the downloading of videos from the playlists and extracted metadata, such as *youtube-video-id*, *youtube-video-title*, *youtube-video-url*, and *video-duration*. These tasks were accomplished using public Python libraries, such as OpenCV, for video data manipulation.

[1]**The MUVY dataset is available at:** https://swperfi-project.github.io/Pages-dev/MuvyDataset-page/

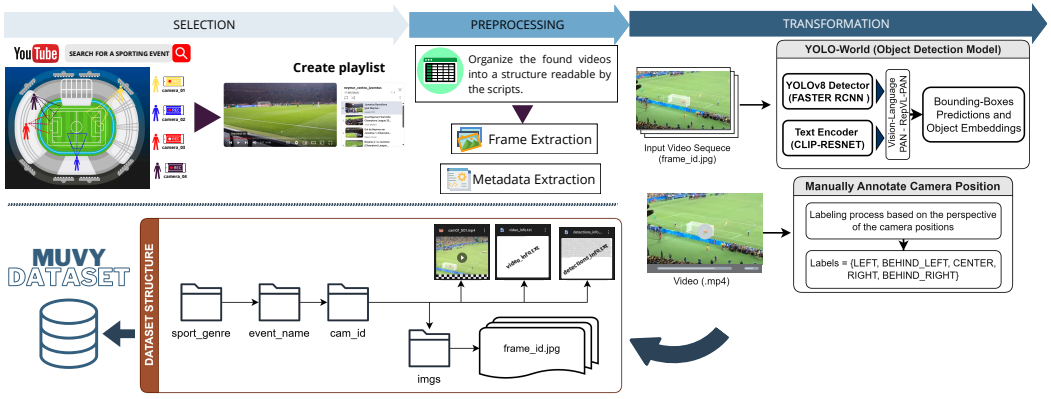

Fig. 1. Dataset Construction Pipeline: searching, collecting, and structuring the extracted sports video content.

(3) **Transformation:** To transform the videos into representative data, two types of annotations are performed: an automated one using object detection and a manual one focusing on the positioning of the cameras recording the videos. These annotations will be explained in Section 5.
(4) **Structuring Data:** The downloaded videos (*.mp4*) and their respective metadata (*.txt*) were organized in a hierarchical structure containing types of sports, events, and sources of videos. Within each sport category, the events were organized into subfolders, each representing different viewpoints. Each folder contained the raw video file and its respective metadata.

The final stages of the KDD process, Interpretation and Evaluation, involved using structured data to develop and test models. This work focuses on constructing and expanding the dataset, as platforms like YouTube continuously update with new content. The dataset was evaluated for its impact and potential applications. Preliminary experiments, such as camera pose estimation using CNNs, showed promise in classifying camera positions in soccer matches. Real-world tests, like selecting the closest camera based on object detection, explored the dataset's utility for video understanding tasks. The evaluation highlighted the dataset's strengths and areas for future research, detailed in Section 6.

5. DATASET CHARACTERISTICS

By the time this article was finished, the created dataset contains 141 videos recorded during 31 sports events. The selected videos present varied and unconventional perspectives compared to official broadcast videos, presenting a variety of angles and camera stability due to hand-held recording, as well as diverse visual quality that depends on the resolution of the device used for recording. Table II provides a summary of the dataset. The following highlights some key characteristics of the dataset:

Table II. Dataset summary.

Sport Genre	Events	Videos	Frames	Duration
Athletics	8	50	174,388	1:46:33
Soccer	8	35	70,774	0:39:48
Basketball	5	13	14,105	0:07:53
American Football	4	30	44,692	0:24:55
Artistic Gymnastic	3	6	18,841	0:10:28
Tennis	2	5	7,064	0:03:55
Cricket	1	2	3,316	0:01:51
TOTALS	**31**	**141**	**333,180**	**3:15:23**

Multi-Perspective Views. The dataset contains video captured simultaneously from multiple viewpoints within the sports venue. So, in order to maintain the idea of multiple viewpoints and

different perspectives, videos that captured the same action simultaneously from various different angles were selected. The idea of the viewpoint distribution across the environment is shown in the top left corner of Figure 1.

Diverse Sporting Events. In order to represent a diverse range of scenarios, movements, interactions, and perspectives typically observed in sport matches, the proposed dataset encompasses a variety of 7 different sporting genre (i.e. football/soccer, american football, artistic gymnastics, athletics, basketball, tennis, cricket). The dataset has videos covering different environmental scenarios, offering a wide range of movement patterns, interactions between players and environmental factors. See Table II for more details. **Duration.** There is a total of 3:15:23 hours of recorded videos in the dataset. The original duration length of each video was kept the same as the YouTube published version. The cumulative distributions of the videos duration for each sport genre are shown in the Table II.

Camera Annotation. The videos in the dataset were labeled based on camera positions, categorized as *LEFT, BEHIND_LEFT, CENTER, RIGHT and BEHIND_RIGHT* (Figure 2). These labels were defined based on the most common angles and views in sports with a rectangular field, such as soccer and tennis, among other sports present in MUVY. The idea of adding these labels was inspired by the SoccerNet-V2 [Deliege et al. 2021], which includes camera label annotations. Obtaining such annotations often requires location sensors, making this a notable advantage over other multiview datasets. These camera annotations are manually performed by an observer watching the videos and labeling the camera positions.

Object Detection Annotation. In the *detection_info.txt*, each line corresponds to one classified object in the associated frame_id. The format for each line is *<frame_id> <object_class_name> <x1> <y1> <x2> <y2>*, where the four values x1,y1,x2,y2 are coordinates representing the bounding box around the detect object. To annotate the objects in the video frames, YOLO-World was used to processes entire images in a single pass, predicting bounding boxes and its coordinates. The initial automated detection is followed by manual validation to ensure accurate and comprehensive annotations. YOLO-World is an advanced real-time object detection model with open-vocabulary capabilities. As shown in Figure 1, it consists of three main components: (1) a YOLO detector that uses a CNN to extract multi-scale features from images, (2) a CLIP text encoder that converts text descriptions into embeddings, and (3) a custom network (RepVL-PAN) for cross-modality fusion of image features and text embeddings [Cheng et al. 2024].

Therefore, there are significant variations in the recorded scenarios in terms of camera motion, stability, orientation, viewpoint, number of players in action on the field, angle and distance diversity, and backgrounds. These variations create a challenging dataset for multi-camera video analysis tasks and applications, such as multi-camera object tracking, camera pose estimation, best view selection, and others.

6. POTENTIAL APPLICATIONS

The dataset creation approach enables various applications in computer vision, particularly in video understanding. Key areas include camera pose estimation and identifying the nearest camera based on object detection, especially useful in sports contexts.

6.1 Camera Pose Estimation

The dataset can be used to validate and test camera pose estimation algorithms, ensuring that they work in real world scenarios and with all the challenges previously discussed in this work with regard to user-recorded videos. Additionally, it allows machine learning models to train to identify and infer camera positions by leveraging the various perspectives of the same event. To explore this capability,

an initial experiment was conducted using a CNN model. The experiment involved taking a sample from a soccer event, which included 5 videos (i.e., 5 cameras). The videos were synchronized, and the extracted frames, along with the associated camera labels, were used to train a model for camera position classification in soccer matches. The model consists of four convolutional blocks, with a max pooling layer between these blocks. Following this, there is a fully connected network with 512 neurons, all using the *ReLU* activation function. The model will produce class probabilities for five classes using *softmax*. In Figure 2, it is possible to see in a) the reference that illustrates the positioning of each view, followed by five samples of the annotated videos; and also in b) some external inputs from images found on the internet, which were tested with the classification model.

Fig. 2. a) Camera perspectives labeling for a sports field divided into five positions and the sample of their views. b) Model's predictions are indicated at the top of each image. Green check: correct, Red X: incorrect prediction.

The model correctly identified some of them, but also produced some false positives. The model evaluation was conducted in terms of accuracy and loss values during training and validation. The training accuracy was 81.37%, while the validation accuracy was 71.15%, indicating that the model generalizes relatively well to unseen data, but there is still room for improvement. Additionally, the losses observed during training and validation reflect this performance, suggesting that adjustments to the model or the use of transfer learning with pre-trained models may be necessary to improve overall accuracy. This was an initial experiment, and further investigation is needed to improve the camera position classification model. However, the dataset can be utilized effectively in these scenarios.

6.2 Closest Camera Selection based on Object Detection

Object detection models can be used to select the closest camera from among multiple videos that have simultaneously captured the same action from different viewpoints. Figure 3 illustrates the general steps involved in the object detection-based closest camera selection process. Initially, the input videos of the *amerFootball event 03* are synchronized in time, ensuring that each frame corresponds to the same moment of action across all five videos. Next, the object detection model, YOLO-World [Cheng et al. 2024], is used to detect the bounding-boxes of the objects of interest (i.e., 'player', 'goalkeeper', 'referee', 'sports ball') within each frame. The goal is to determine the area occupied by each detected object in the frame. The camera providing the closest view is identified as the one capturing the largest relative area of the frame occupied by the detected objects, that is, assuming the closest camera as the one with the highest percentage of image area occupied by the detected objects. The relative area occupied by detections in a frame can be calculated using the formula: Frame ID occupied relative area $= \frac{\sum \text{detection.area}}{\text{Frame ID area}}$, where Frame ID area = img_height × img_width and detection.area = bounding_box_height × bounding_box_width. A larger object area indicates a closer and more detailed viewpoint of the current action.

Therefore, the dataset proposed in this work can be used to validate and test the closest camera selection algorithm based on object detection, evaluating its functionality in real-world scenarios. Additionally, this dataset can be used to train and evaluate models specifically designed to detect objects of interest in sports scenes. Beyond annotated object characteristics, the dataset allows for the extraction of video quality metrics that can influence the selection of the best viewpoint. Combining

Fig. 3. Camera Selection Overview: (1) Synchronize videos; (2) Detect objects and calculate areas using YOLO-World; (3) Select the camera with the largest occupied area.

these metrics with the object detection results can improve the final results of the closest camera selection method. Furthermore, person re-identification (re-id) in videos is a task for applications that require tracking individuals across multiple cameras. For example, the MUVY dataset can be used to evaluate object tracking models that quickly re-identify players across different camera viewpoints at any given moment in a match. Such a system has numerous applications, including tracking players across multiple cameras to create automatic highlight videos that focus on a single player.

7. CONCLUDING REMARKS

The MUVY dataset includes a total of 31 events, 141 videos and a combined duration of over 3 hours of footage across multiple sports. Despite its limitations, such as the size of the dataset, which may seem small for certain applications, the dataset has great potential for contribution. Compared to related works, it already shows interesting numbers. The data source, YouTube, tends to grow continuously, allowing the dataset to expand. The manual effort to annotate camera positions and validate identified objects is another limitation that requires investigation to reduce these efforts. The quality of user-generated videos can vary, presenting shakes, obstructions, and low resolution, which is reflected in the extracted frames. This is part of the challenging context proposed, differentiating it from available multi-camera datasets, which are generally more controlled. The MUVY dataset offers a valuable resource for tasks related to video understanding, particularly in camera pose estimation and optimal viewpoint selection for sports events. Initial experiments, such as the identification of camera positions based on frame regions and the selection of the closest camera to game action using object detection to extract relevant features from the video and areas occupied by key objects, have shown promising results. Therefore, this dataset has the potential to contribute significantly to future research in computer vision.

Acknowledgment

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, 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

Bailer, W., Pike, C., Bauwens, R., Grandl, R., Matton, M., and Thaler, M. Multi-sensor concert recording dataset including professional and user-generated content. In *Proceedings of the 6th ACM multimedia systems conference*. pp. 201–206, 2015.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., and Shan, Y. Yolo-world: Real-time open-vocabulary object detection. *arXiv preprint arXiv:2401.17270*, 2024.

Cho, B., Le, B. M., Kim, J., Woo, S., Tariq, S., Abuadbba, A., and Moore, K. Towards understanding of deepfake videos in the wild. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. pp. 4530–4537, 2023.

Citraro, L., Márquez-Neila, P., Savare, S., Jayaram, V., Dubout, C., Renaut, F., Hasfura, A., Ben Shitrit, H., and Fua, P. Real-time camera pose estimation for sports fields. *Machine Vision and Applications* vol. 31, pp. 1–13, 2020.

Cricri, F., Roininen, M., Mate, S., Leppänen, J., Curcio, I. D., and Gabbouj, M. Multi-sensor fusion for sport genre classification of user generated mobile videos. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 1–6, 2013.

de Aguiar Salvi, A. and Barros, R. C. An experimental analysis of model compression techniques for object detection. *Proceedings of the 8th KDMiLe, 2020, Brasil.*, 2020.

Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Dueholm, J. V., Nasrollahi, K., Ghanem, B., Moeslund, T. B., and Van Droogenbroeck, M. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4508–4519, 2021.

Fayyad, U. Knowledge discovery in databases: An overview. In *International Conference on Inductive Logic Programming*. Springer, pp. 1–16, 1997.

Giancola, S., Amine, M., Dghaily, T., and Ghanem, B. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. pp. 1711–1721, 2018.

Gonçalves, L. A., Zampolo, R. F., and Barros, F. B. A multi-stream dense network with different receptive fields to assess visual quality. In *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*. SBC, pp. 65–72, 2019.

Held, J., Cioppa, A., Giancola, S., Hamdi, A., Ghanem, B., and Van Droogenbroeck, M. Vars: Video assistant referee system for automated soccer decision making from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5085–5096, 2023.

Li, Y., Chen, L., He, R., Wang, Z., Wu, G., and Wang, L. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 13536–13545, 2021.

Naab, T. K. and Sehl, A. Studies of user-generated content: A systematic review. *Journalism* 18 (10): 1256–1273, 2017.

Olagoke, A. S., Ibrahim, H., and Teoh, S. S. Literature survey on multi-camera system and its application. *IEEE Access* vol. 8, pp. 172892–172922, 2020.

Perera, A. G., Law, Y. W., Ogunwa, T. T., and Chahl, J. A multiviewpoint outdoor dataset for human action recognition. *IEEE Transactions on Human-Machine Systems* 50 (5): 405–413, 2020.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*. Springer, pp. 17–35, 2016.

Saini, M., Venkatagiri, S. P., Ooi, W. T., and Chan, M. C. The jiku mobile video dataset. In *Proceedings of the 4th ACM multimedia systems conference*. pp. 108–113, 2013.

Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., and Wang, O. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1279–1288, 2017.

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023.

Wang, M., Shi, D., Guan, N., Yi, W., Zhang, T., and Fan, Z. Multi-target multi-camera tracking with human body part semantic features. In *CIKM*. pp. 199–208, 2019.

Zhang, Y., Bai, Y., Chang, J., Zang, X., Lu, S., Lu, J., Feng, F., Niu, Y., and Song, Y. Leveraging watch-time feedback for short-video recommendations: A causal labeling framework. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. pp. 4952–4959, 2023.

Zhao, Z., Chai, W., Hao, S., Hu, W., Wang, G., Cao, S., Song, M., Hwang, J.-N., and Wang, G. A survey of deep learning in sports applications: Perception, comprehension, and decision. *arXiv preprint arXiv:2307.03353*, 2023.