ITI-CERTH participation in ActEV and AVS Tracks of TRECVID 2024

Konstantinos Gkountakos, Damianos Galanopoulos, Antonios Leventakis, Georgios Tsionkis, Klearchos Stavrothanasopoulos, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

Information Technologies Institute, Centre for Research and Technology Hellas,
6th Km. Charilaou - Thermi Road, 57001 Thermi-Thessaloniki, Greece
{gountakos, dgalanop, aleventakis, tsiogeorgios, klearchos_stav, kioannid, stefanos, bmezaris, ikom}@iti.gr

Abstract

This report presents the overview of the runs related to Ad-hoc Video Search (AVS) and Activities in Extended Video (ActEV) tasks on behalf of the ITI-CERTH team. Our participation in the AVS task involves a collection of five cross-modal deep network architectures and numerous pre-trained models, which are used to calculate the similarities between video shots and queries. These calculated similarities serve as input to a trainable neural network that effectively combines them. During the retrieval stage, we also introduce a normalization step that utilizes both the current and previous AVS queries for revising the combined video shot-query similarities. For the ActEV task, we adapt our framework to support a rule-based classification to overcome the challenges of detecting and recognizing activities in a multi-label manner while experimenting with two separate activity classifiers.

1 Introduction

In this paper, the work carried out in the context of TRECVID 2024 by the ITI-CERTH[1] team in the area of video analysis, retrieval and understanding is presented. ITI-CERTH has participated in TRECVID [1] for many years as it is one of the most popular video understanding challenges. Especially, ITI-CERTH has participated in Search and Semantic Indexing (SIN) tasks under the research network COST292 (TRECVID 2006-2008) and the MESH and K-SPACE (TRECVID 2007-2008) EU-Funded research projects, correspondingly. From 2009 to 2015 [2, 3, 4, 5, 6, 7, 8] ITI-CERTH team has participated as a stand-alone organization in a significant number of tasks including but not limited to SIN, KIS, INS, and MED. In both 2016 [9] and 2017 [10], ITI-CERTH participated in the AVS, MED, INS and SED tasks. In 2018 [11], ITI-CERTH participated in the AVS, INS and ActEV; in 2019 [12], the participation was limited to the ActEV task. In 2020 [13] ITI-CERTH participated in the AVS, DSDI and ActEV tasks; in 2021 [14] and 2022 [15] ITI-CERTH participated in the AVS and ActEV tasks. Lastly, in 2023 [16] the participation was only for the AVS task. Considering the above-mentioned submissions, we aim to evaluate improved algorithms and systems. This year, ITI-CERTH participated again in AVS and ActEV tasks. The following sections will present the employed algorithms and the evaluation of the runs during the AVS and ActEV tasks, respectively.

[1]Information Technologies Institute - Centre for Research and Technology Hellas

2 Ad-hoc Video Search

The TRECVID 2024 [17] Ad-hoc Video Search (AVS) task aims to develop a system for retrieving a ranked list of 1000 video shots for each ad-hoc textual query, ranked from the most relevant to the least relevant shot for the query. To address this task, we employed a new approach by focusing on utilizing a diverse set of pre-trained image-text models to improve AVS task performance. Specifically, we utilized several cross-modal networks in our pipeline, pushing the boundaries of model combination. Based on previous methodologies, such as those employed in [18], where fixed model weights were a crucial factor, our approach focuses on the dynamic interaction between these pre-trained models. We recognized that each model captures unique aspects of the multimodal relationship, but the weightage of these features varies depending on the input data. Therefore, rather than relying on fixed weights, we designed and trained a network that learns to weigh each model's contributions optimally. This allows our system to adaptively shift focus toward the most effective models for any given input, enhancing accuracy and performance in real-world scenarios.

2.1 Approach

In our participation in AVS 2024, inspired by [18], we employ a collection of $N = 5$ cross-modal network families, each leveraging multiple pre-trained models to achieve robust video-text matching. Specifically, we use the text-image network families: CLIP [19], BLIP [20], BLIP-2 [21], SLIP [22], and BEiT3 [23]. For each network family, we apply multiple pre-trained model variants (as listed in Table 1) to compute feature representations for the D video shots that form our dataset (a shot is denoted as v_d where $d = \{1, 2, \ldots, D\}$). Likewise, each of the Q queries, s_q, is processed through these models to extract corresponding query features.

Based on these features, for each pre-trained model, we calculate the cosine similarities $sim^{i,n}(s_q, v_d)$ between all video shots and a query s_q, where i denotes a pre-trained model within a model family (e.g. CLIP, BLIP, etc) and n identifies the network family. To combine the similarities of the models within the same network family, for a given video shot-query pair (s_q, v_d), we summarize the similarity scores as follows: $sim^n(s_q, v_d) = \sum_i sim^{i,n}(s_q, v_d)$.

In order to combine all these similarities $\mathbf{sim}(s_q, v_d) = [sim^1(s_q, v_d), sim^2(s_q, v_d), \ldots, sim^N(s_q, v_d)]$ and to calculate the final similarity y for each query-video shot pair, we designed a trainable neural network. Instead of averaging these five similarities, our network learns which of them is more of less useful since each model processes this data differently, highlighting various relationships between visual and textual features. This network $\mathcal{G}(\cdot)$ is composed of two fully connected (FC) layers (comprising 64 and 32 nodes, respectively), followed by ReLU activation functions for non-linearity. This architecture dynamically evaluates the importance of each model output, effectively creating a weighted fusion system that can adapt to the complexities of various queries.

The output of the neural network, y, is given as follows:

$$y = \mathcal{G}(\mathbf{sim}(s_q, v_d)) = W^{(2)} \cdot \text{ReLU}\left(W^{(1)}\mathbf{sim}(s_q, v_d) + \mathbf{b}^{(1)}\right) + b^{(2)}$$

where $W^{(1)}$, $W^{(2)}$, $\mathbf{b}^{(1)}$ and $b^{(2)}$ are the trainable hyper-parameters.

To further improve the query-video shot similarities we introduced a similarity normalization procedure across queries utilizing this year's and older AVS queries as background queries bq_j, where $j = \{1, 2, \ldots, J\}$ and J is the number of the background queries. More specifically, at the retrieval stage, for a given query-video shot pair (s_q, v_d) and a specific network family n, besides the similarity $sim^n(s_q, v_d)$ we also calculate the similarities between the shot v_d and all the background queries bq_j forming a similarity vector $\mathbf{sim}^n_{bg}(s_q, v_d) = [sim^n(s_q, v_d), sim^n(bq_1, v_d), \ldots, sim^n(bq_J, v_d)]$, and then we normalize this vector using $l2$ normalization:

$$\hat{\mathbf{sim}}^n_{bg}(s_q, v_d) = \frac{\mathbf{sim}^n_{bg}(s_q, v_d)}{\|\mathbf{sim}^n_{bg}(s_q, v_d)\|_2}$$

resulting in a new revised vector $\hat{\mathbf{sim}}^n_{bg}(s_q, v_d) = [\hat{sim}^n(s_q, v_d), \hat{sim}^n(bq_1, v_d), \ldots, \hat{sim}^n(bq_J, v_d)]$.

Following this normalization across all N network families, the revised similarities that are used as input to our trained network $\mathcal{G}(\cdot)$ are formed as follows:

$$\hat{\mathbf{sim}}(s_q, v_d) = [\hat{sim}^1(s_q, v_d), \hat{sim}^2(s_q, v_d), \ldots, \hat{sim}^N(s_q, v_d)]$$

.

Table 1: The cross-modal network families and their pre-trained model variations utilized in our AVS 2024 participation.

Network Family	Pre-trained Model	Network Family	Pre-trained Model
BLIP	base_coco	BEiT3	base_coco (retrieval)
	base_flickr		base_flickr30k (retrieval)
	large_coco		large_coco (retrieval)
	large_flickr		large_flickr30k (retrieval)
BLIP 2	itm_coco	CLIP	RN101
	itm_pretrain		RN50
	itm_pretrain_vitL		RN50x4
SLIP	base_CC12M		RN50x16
	base_CC3M		RN50x64
	base		ViT-B/16
	large		ViT-B/32
	small		ViT-L/14

2.2 Submission

Our network $\mathcal{G}(\cdot)$ is trained using a combination of four large-scale video captioning datasets: MSR-VTTT [24], TGIF [25], ActivityNet [26] and Vatex [27]. The V3C2 [28] dataset is utilized to evaluate the network's performance. Moreover, we examine the performance of our runs on the V3C2 datasets for the queries of years 2022-2023. The evaluation measure we use is the mean extended inferred average precision (MxinfAP).

This year, we submitted three runs for the AVS 2024 main task and three additional runs for the AVS progress subtask. Overall, we evaluate our methods on 40 ad-hoc queries (20 from the main task and 20 from the progress subtask). The submitted runs are briefly described below:

- ITI_CERTH.24_run_1: Network $\mathcal{G}(\cdot)$ is trained to combine text and video similarities from various cross-modal networks. The input similarities have been normalized, considering the 2022, 2023, and 2024 queries as background queries.
- ITI_CERTH.24_run_2: Similar to run 1, but the input similarities have been normalized, considering only 2024 queries as background queries.
- ITI_CERTH.24_run_3: Network $\mathcal{G}(\cdot)$ of runs 1 and 2 without the normalization step in the input similarities.

2.3 Experimental Results

This section presents the results of the Main and Progress tasks. Table 2 presents the official results of our submissions for the main AVS task, as well as the results of our internal evaluation of AVS 2022 and 2023 queries. The ITI_CERTH.24 run_1 and the ITI_CERTH.24 run_2 runs, where we utilize background queries at the similarity normalization step, constantly outperform by a significant margin the ITI_CERTH.24 run_3 where no normalization step was performed. Similar results were observed in our internal experiments on the AVS 2022 and AVS 2023 queries, as shown in Table 2, where run 3 achieved significantly lower performance than runs 1 and 2.

Similarly, Table 3 summarizes the evaluation results of our runs for the Progress AVS task on sets A and B. The ITI_CERTH.24 run_2, where only the AVS 2024 queries were used as background

Table 2: Mean Extended Inferred Average Precision (MXinfAP) for all submitted runs for the 2022, 2023 and 2024 fully automatic AVS tasks.

Run id:	2022	2023	2024
ITI_CERTH.24 run_1	0.278	**0.335**	**0.360**
ITI_CERTH.24 run_2	**0.279**	0.328	0.353
ITI_CERTH.24 run_3	0.246	0.269	0.273

Table 3: Mean Extended Inferred Average Precision (MXinfAP) for all submitted runs for the fully automatic AVS Progress task.

Run id:	Progress Set A	Progress Set B
ITI_CERTH.24 run_1	0.267	0.358
ITI_CERTH.24 run_2	**0.268**	**0.361**
ITI_CERTH.24 run_3	0.235	0.265

queries in the similarity normalization step, outperforms by a small margin the ITI_CERTH.24 run_1, where we utilize queries from three years. However, both runs outperform the ITI_CERTH.24 run_3, where no normalization step was applied to the similarities.

The results from the above tables highlight the merit of the normalization step, which consistently improves performance regardless of the query list used.

Figure 1 shows the performance of our submitted runs in the AVS 2024 main task compared to all submitted runs from the other teams. Moreover, Figures 2 and 3 show the performance and evolution of all submitted runs in the AVS progress task from 2022 to 2024. Both figures demonstrate that our methods have consistently improved each year and achieved competitive results.

3 Activities in Extended Video

In activity recognition systems, multimedia data captured from cameras in diverse indoor and outdoor environments is analyzed to identify activities of depicted objects. This field has gained significant attention due to its practical applications in areas such as surveillance and traffic monitoring. However, effective activity recognition in these systems faces several challenges: video resources are often untrimmed, multiple activities can occur simultaneously, and interactions between objects can be complex. These factors, along with the need for real-time analysis, make manual processing and interpretation impractical and highlight the importance of automated methods.

Towards this goal, the Activities in Extended Videos (ActEV) challenge promotes the research and development of real-time activity detection methods in surveillance scenarios. In our approach, the task of activity recognition is addressed through a sequence of distinct steps, beginning with object detection and tracking followed by the classification of the activities. We utilized the YOLOv8 model [29] in conjunction with the BoT-SORT [30] tracking algorithm, to detect and track key Objects of Interest (OoI), including persons, vehicles, bags, laptops, and cellphones. For activity classification, we segmented activities into three categories: Person-Related (PR), Vehicle-Related (VR), and Person-Vehicle-Related (PVR) activities. PR activities involve persons not overlapping with vehicles, including also bags, laptops, and cellphones, and are classified using a rule-based method customized to each activity type. PVR and VR activities, involving overlapping persons and vehicles as well as standalone vehicles, are classified with a 3D-ResNet-based classifier [31]. The MEVA dataset [32] was used to train and validate the activity classifier using the official Kitware annotations[1]. All 20 activity classes are depicted in Table 4.

[1]https://gitlab.kitware.com/meva/meva-data-repo/-/tree/master/annotation/DIVA-phase-2/MEVA

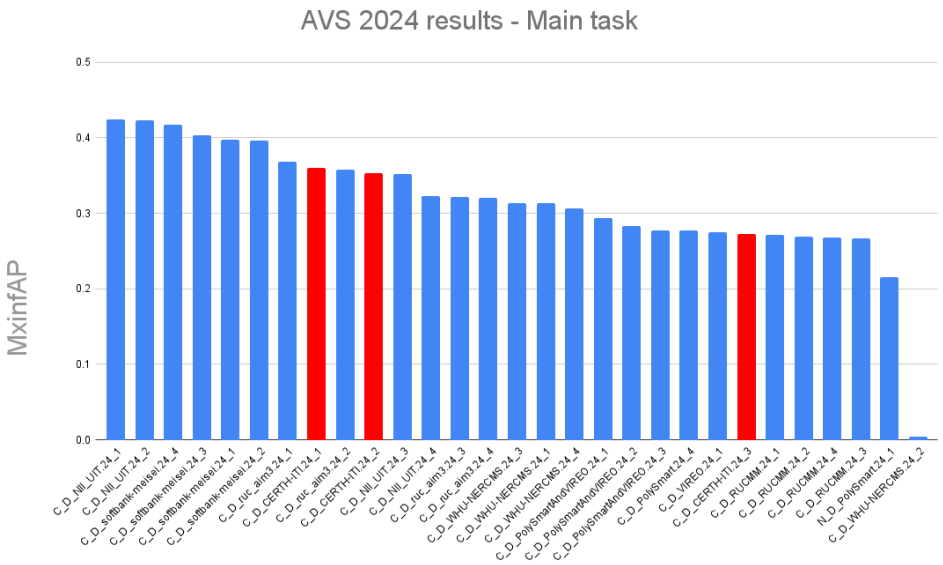

Figure 1: AVS 2024 ranking of all submitted runs regarding the Main task, according to MXinfAP. Red bars indicate our submitted runs.

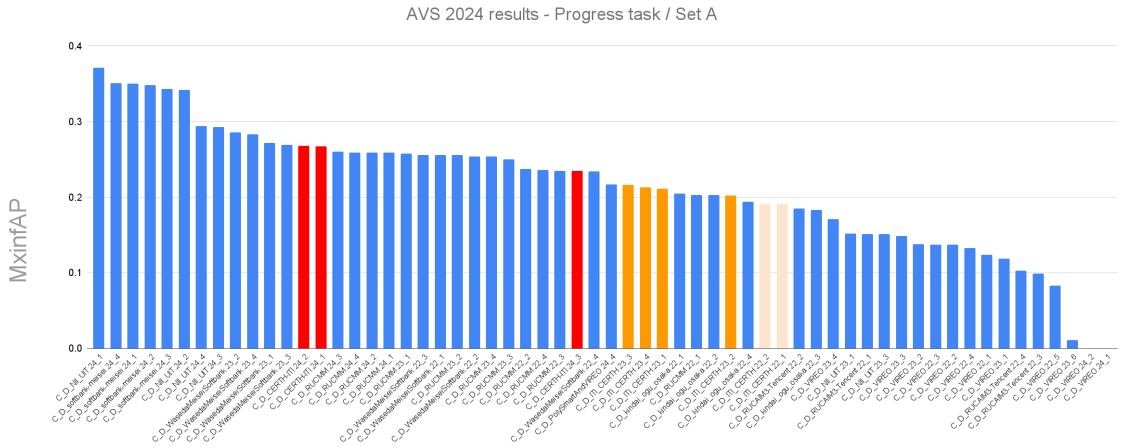

Figure 2: AVS 2024 ranking of all submitted runs regarding Set A of the Progress task, according to MXinfAP. Red, orange, and pink bars indicate our submitted runs for the years 2024, 2023, and 2022 respectively.

3.1 Approach

In this work, the task of activity recognition and localization is addressed across a set of videos $V = \{v_i\}$ to identify a set of activities $A = \{a_i\}$ for all the Objects of Interest. Each activity a_i is described by a type t_i according to a set of predefined classes and a temporal location $l = (t_{start}, t_{end})$ that indicates the start and end of it within a given video.

Considering the challenges of detecting activities in surveillance videos, this work focuses on effectively detecting and tracking the OoI, and identifying their activities. Building on our previous submissions [13, 14, 15] in the activity detection task of the ActEV challenge, this year we have introduced three significant enhancements: (1) the adoption of an updated version of YOLO series, specifically the YOLOv8 [29] model, for improved detection accuracy, (2) the integration of the BoT-SORT [30] algorithm for enhanced tracking performance, and (3) the implementation of a new

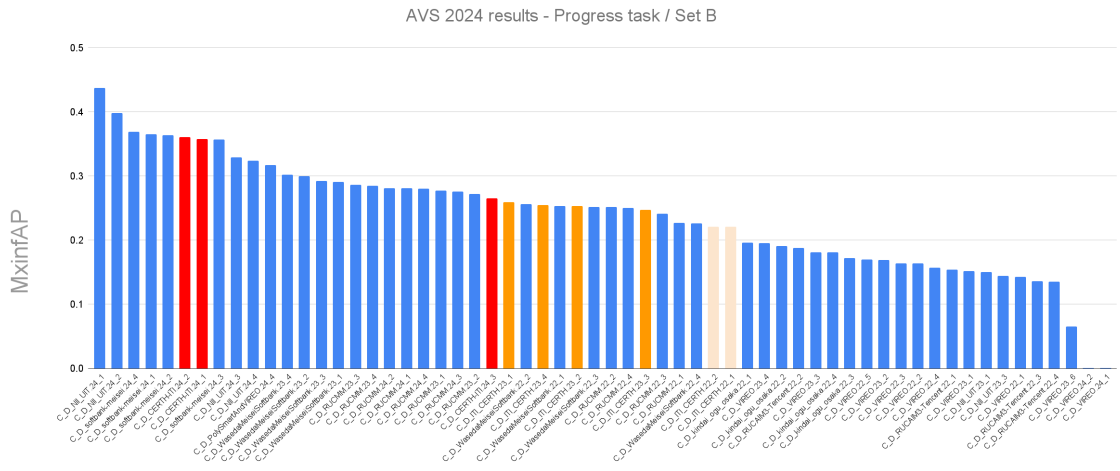

Figure 3: AVS 2024 ranking of all submitted runs regarding Set B of the Progress task, according to MXinfAP. Red, orange, and pink bars indicate our submitted runs for the years 2024, 2023, and 2022 respectively.

Table 4: Activity classes in ActEV challenge 2024, divided into Person-Related (PR), Vehicle-Related (VR) and Person-Vehicle-Related (PVR) classes.

Activity Classes	
PR Classes	**VR & PVR Classes**
person_reads_document	person_closes_vehicle_door
person_enters_scene_through_structure	person_enters_vehicle
person_exits_scene_through_structure	person_exits_vehicle
person_stands_up	person_opens_vehicle_door
person_sits_down	vehicle_starts
person_talks_to_person	vehicle_stops
person_picks_up_object	vehicle_turns_left
person_puts_down_object	vehicle_turns_right
person_opens_facility_door	
person_texts_on_phone	
person_interacts_with_laptop	
person_transfers_object	

rule-based classification method is introduced for PR activities, alongside our established classifier for VR and PVR classes.

3.1.1 Object Detection and Tracking

The first step involves object detection and tracking, conducted frame-by-frame across each video to identify and track all objects for subsequent activity analysis. The largest and most accurate model of the YOLOv8 [29] series, YOLOv8x, is incorporated as the object detector in our pipeline, offering cutting-edge performance. For the experiments, we employed a pre-trained YOLOv8x model, which was trained on the Microsoft Common Objects in Context (MS COCO) [33] dataset, to detect objects classified into the following categories: "person", "vehicle" (depicting the classes "car", "motorcycle", "bus", and "truck"), "laptop", "cellphone" and "bags" (representing the classes "backpack", "handbag" and "suitcase"). YOLOV8x achieves 53.9% Average Precision (AP) on MS COCO, running at 274 FPS on an A100 TensorRT GPU.

Apart from object detection, Ultralytics repository [2] also provides advanced tracking algorithms

[2]https://docs.ultralytics.com/tasks/

that not only identify the location and class of objects within a frame but also maintains a unique ID for each detected object as the video progresses. BoT-SORT algorithm [30] has been utilized as our tracker, which addresses the limitations of prior SORT-like trackers, such as Simple Online Realtime Tracking (SORT) [34], Deep Simple Online Realtime Tracking (DeepSort) [35] and Joint Detection and Embedding (JDE) [36] by incorporating features from the novel ByteTrack [37] algorithm. BoT-SORT leverages both motion and appearance information, incorporates camera-motion compensation, and employs an enhanced Kalman filter state vector for improved box localization and robust detection-to-tracklet associations. The output of this process is a set of tracked objects, $O = \{o_i\}$, where each object o_i is defined by its bounding boxes across all video frames in which it was tracked, $o_i = \{(x_{left}, y_{top}, width, height)_{t_1}, ...\}$.

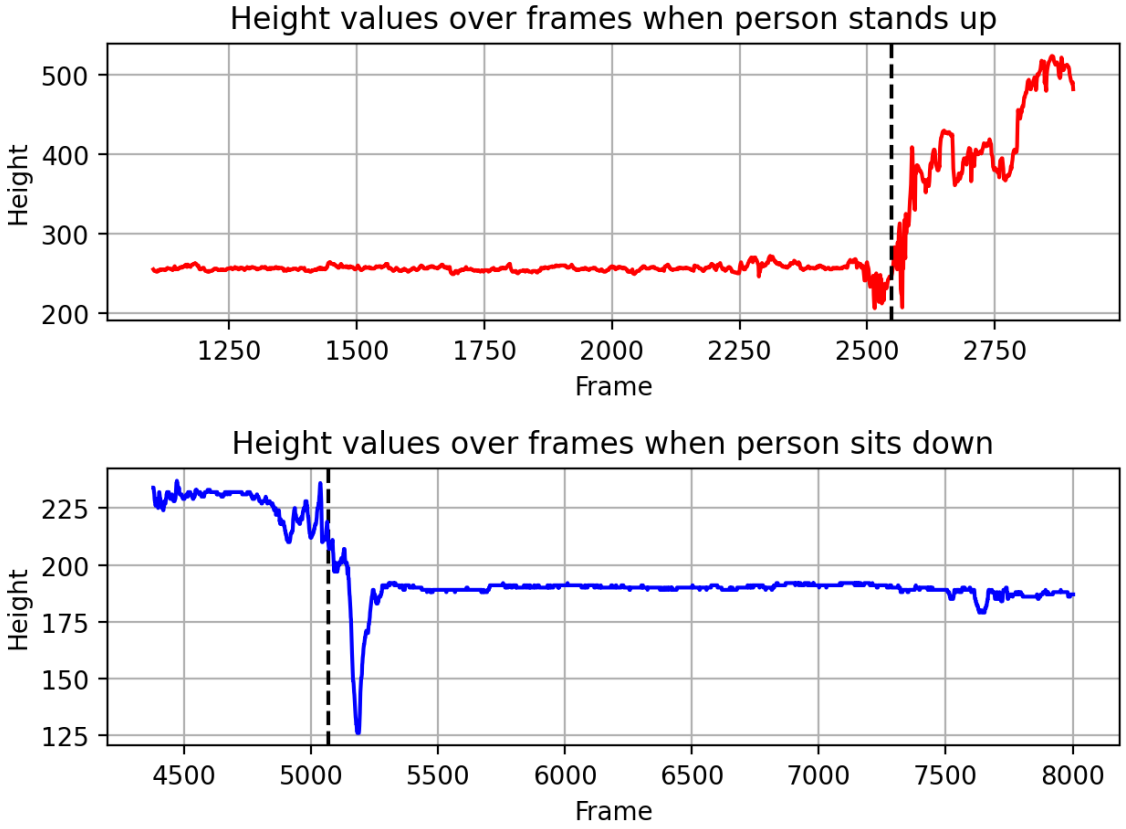

Figure 4: Bounding box height of person objects across frames. On the top is depicted an example of a person standing up and on the bottom when a person sitting down. The black dashed line highlights the change point of the height values; indicating the beginning of the corresponding activity.

3.1.2 PR Activity Recognition

For PR activities, we propose a rule-based classification method leveraging the distinct characteristics of each activity:

- Person-Laptop and Person-Phone Interactions: The overlap of a person's bounding box with a laptop serves as an indicator of the activity *person_interacts_with_laptop*. However, to minimize the false positives, such as when a person in the scene passes by the laptop, we add a duration threshold for the overlap of 150 frames. Only if the overlap persists beyond this threshold we label the activity as *person_interacts_with_laptop*. The same logic applies to *person_texts_on_phone*, though here the overlap is nearly 100% as the person holds the phone, resulting in close bounding box alignment.

- Sitting and Standing Actions: The decrease in a person's bounding box height serves as a key indicator of the activity *person_sits_down*. However, temporary height reductions, such as when a person overlaps with another object, can produce similar results. The key difference lies in the duration of the change. When a person sits down, the decrease in bounding box height is sustained, whereas during a temporary occlusion, the height decreases but quickly recovers as the person reappears from the overlapped object. A similar pattern is observed in activity *person_stands_up*, where in that situation the height increases. The above observation is illustrated in Figure 4.
- Person-to-Person Interactions: To identify closely related persons that can be involved in the activity *person_talks_to_person*, the distance between the centres of the corresponding bounding boxes is calculated. Based on distances observed for this activity in the MEVA dataset [32] training and validation sets, as shown in Figure 5, we establish the mean distance threshold of 75 pixels. Additionally, to avoid including people who are just passing by each other briefly, we introduce a duration restriction. Specifically, if the distance of the persons remains below 75 pixels for more than 150 frames, we classify the activity as *person_talks_to_person*.

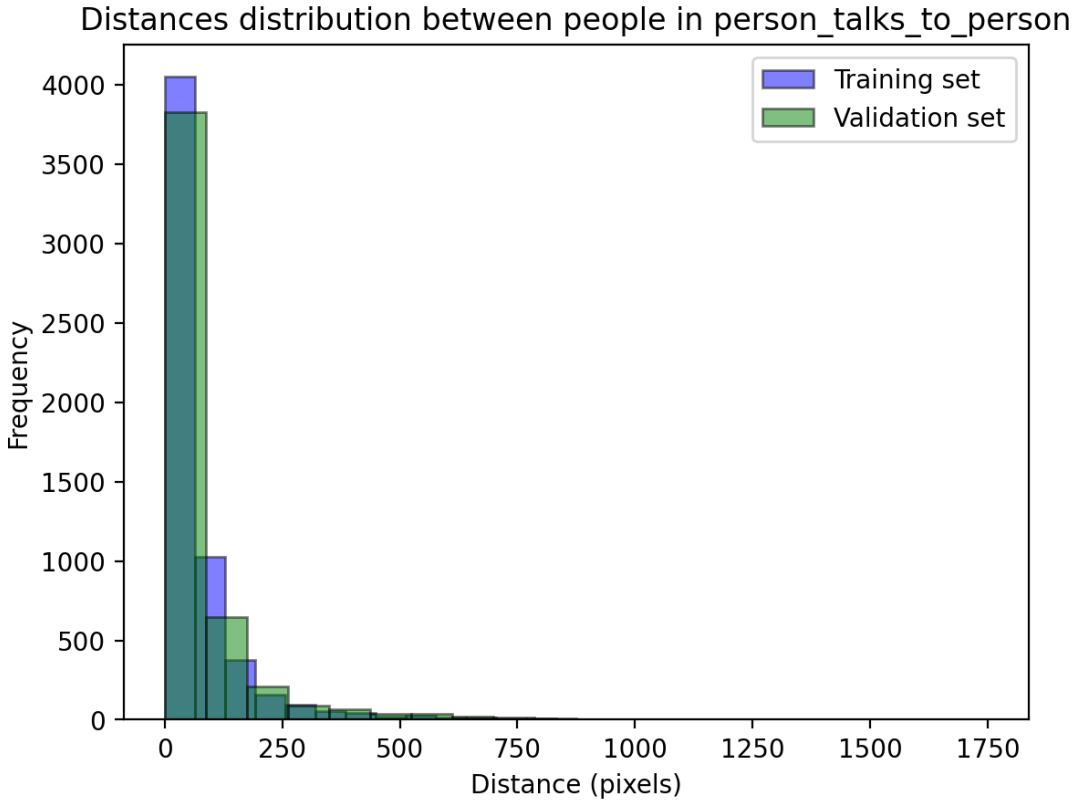

Figure 5: Distribution of distances between people in activity person_talks_to_person of MEVA dataset for both training and validation set.

- Object Transfer Activities: Identifying overlaps between bags and persons could be an indicator for the activity *persons_transfers_object*, after filtering out static bags among frames. Furthermore, if the bag's y-coordinate increases during the overlap, the activity is classified as *person_picks_up_object*. On the other side, if the bag's y-coordinate decreases, the activity is classified as *person_puts_down_object*.

3.1.3 VR and PVR Activities Recognition

The final step of our pipeline is the activity recognition for VR and PVR activities. As a pre-processing step, we remove all the static vehicles from the VR pool, as ActEV challenge activities do not involve

stationary vehicles. For PVR pool, where a person's bounding box overlaps with that of a vehicle, we compute the union of the spatial data before feeding it into the deep learning model. The 3D-ResNet [31] is employed to label the tracked objects, a deep learning architecture that effectively handles spatiotemporal data with 3D convolutional layers. Its architecture consists of four sequential bottleneck blocks, where each block includes three 3D-convolution layers (with varying kernel sizes), along with batch normalization, and ReLU activation layers. For this challenge, we initialized the model with pre-trained weights from the Kinetics [38] dataset and then fine-tuned it in a multi-label manner using the MEVA dataset. The model assigns scored labels to every batch of 16 frames of a detected object's trajectory. To produce more accurate activity proposals and reduce the number of false alarms, a threshold T_{hight} has been established, with activities scoring below this value are less likely to occur.

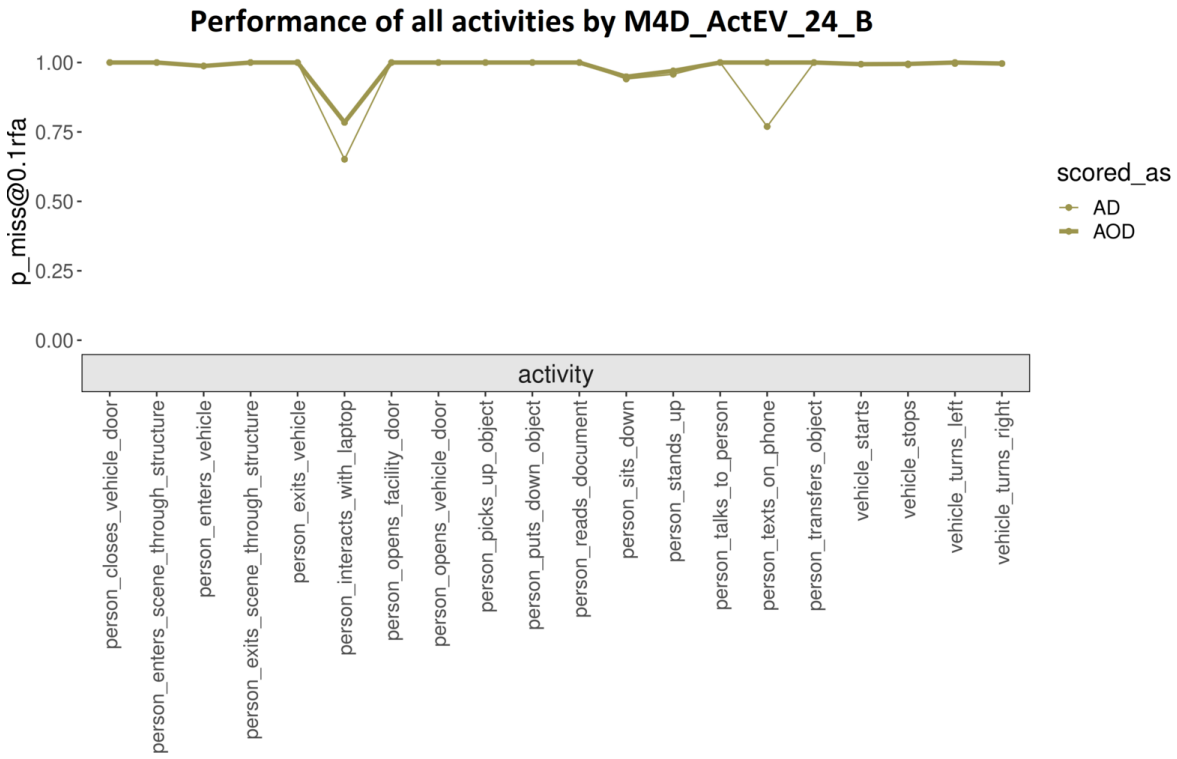

Figure 6: Performance of our system M4D_ActEV_24_B in each activity for Activity and Object Detection (AOD) and Activity Detection (AD).

3.2 Submission

In this section, we present our submitted system, as depicted in Figure 6 and in Table 5 for both Activity and Object Detection (AOD) and Activity Detection (AD):

- M4D_ActEV_24_B : The system was deployed using YOLOv8x [29] for object detection and the BoT-SORT [30] algorithm for tracking, splitting the classes according to Table 4. Our rule-based method was employed for PR activities. The MEVA dataset [32] was utilized to train and validate the activity classifier for VR and PVR activities. Refinement was achieved by adjusting the T_{hight} threshold for the final scores to 75%.

3.3 Experimental Results

In this section, further discussion about the performance of the submitted system is reported. Table 5 presents the results of our system as described in the Submissions section, on both validation and

Table 5: The evaluation results of our M4DSYS_ACTEV_24_B for MEVA validation set and ActEV challenge test set for Activity and Object Detection (AOD) and Activity Detection (AD).

	Validation Set		**Test Set**	
	AOD	**AD**	**AOD**	**AD**
p_miss@0.1rfa	0.9833	0.9697	0.9838	0.9643
nAUDC@0.2rfa	0.9779	0.9640	0.9781	0.9588
Correct Detections	650	1310	-	-
False Detections	60899	60239	-	-
Missed Detections	31355	30695	-	-
Activities	61549	61549	21118	21118

test sets. The validation set results were generated from evaluations that ran locally, while the test set results gained from the public leaderboard[3] of the ActEV challenge. Both sets were evaluated using the "SRL_AOD_V1" and "SRL_AD_V1" scoring protocols for AOD and AD, respectively. To better assess the effectiveness of our new rule-based method for PR activities, Figure 6 illustrates the system's performance across different activities. Notably, our system achieved high accuracy in the *person_interacts_with_laptop* activity, with scores of 0.76 and 0.64 for AOD and AD, respectively, for the p_miss@0.1rfa metric. Additionally, strong performance can be observed in other activities such as *person_texts_on_phone*, *person_sits_down*, and *person_stands_up*.

The success of the rule-based classification method, heavily relies on the object detection of YOLOv8 model. With this in mind, in some cases, misclassifications are observed, such as a detecting newspapers as laptops, or failures to detect objects involved in activities of interest, such as doors (*person_opens_facility_door*, *person_enters_scene_through_ structure*, *person_exits_scene_through_structure*), documents (*person_reads_document*), or items like cups, money or other objects involved in the activities *person_picks_up_object*, *person_puts_down_object*. As a result, while the rule-based method provides a reasonable foundation, it still leads to many false detections and the exclusion of certain activities from our analysis. Another factor impacting performance is camera proximity, where people's appearance relative to the camera affects the spatial information of tracked objects. Future iterations would benefit from calibrating thresholds based on camera type to enhance consistency and reliability.

In VR and PVR activity recognition, the exclusion of static cars has reduced false positives, but many vehicles that follow consistent paths without notable actions still contribute to false detections. As the classifier assigns a label to each 16-frame batch, it can misinterpret stable vehicle trajectories as activity, creating false positives. Additionally, activity labels are currently applied to an entire vehicle trajectory, while annotations may only apply to specific segments (e.g., a *vehicle_turns_left* might occur briefly). Processing full trajectories rather than segments can lead to mislabeling and lower detection accuracy.

While our submission ranked last in ActEV 2024, an overall review of all submissions places this year's entry in 9th position out of 16-a significant improvement over our previous last-place finish in 2022 (Figure 7). This progress, along with strong performance in certain activities and the implementation of new refinement rules, reinforces our confidence for future enhancements and better results.

4 Conclusions

In this paper, the evaluation of ITI-CERTH during the TRECVID 2024 challenge [17] is reported. ITI-CERTH this year participated by developing new techniques and algorithms in the context of AVS and ActEV tasks. In the AVS task, we leveraged several image-text cross-modal network families to enhance our system's performance and we trained a neural network that learns the optimal weighting for each model's contribution, allowing the system to highlight the most informative features for each

[3]https://actev.nist.gov/SRL#tab_leaderboard

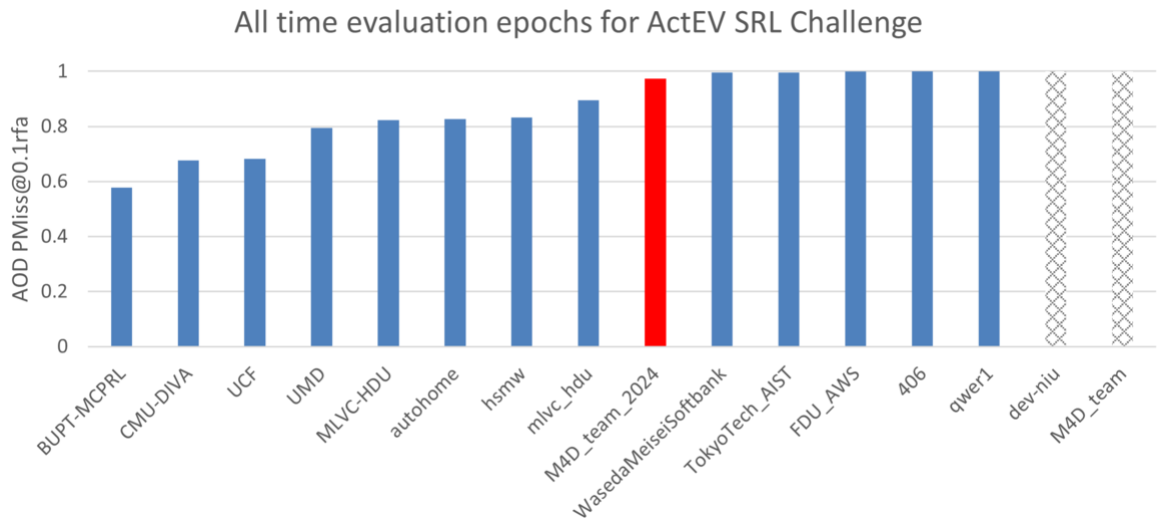

Figure 7: All time submissions for ActEV SRL Challenge based on the primary metric of AOD, Pmiss@0.1RFA. M4D_team and dev-niu metric is not reported.

query. Additionally, a similarity normalization step refines query-video shot similarities, improving overall performance.

Regarding the ActEV task, a three-step pipeline was deployed in order to effectively detect objects, track them and recognize their activities in a multi-label manner enriched by rule-based classification filtering capabilities. The classification of the detected activities is performed spatio-temporally using two separate classifiers; one for the person-related activities and one for the vehicle-related and person-vehicle interaction activities. Though the results are not expected, some aspects of the process seem promising.

5 Acknowledgements

This work was partially supported by the projects PRECRISIS (ISF-101100539), SAFEGUARD (ISF-6006936), funded by the European Union's Internal Security Fund, and AI4TRUST (HE-101070190), funded by the European Commission's Horizon Europe program.

References

[1] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[2] A. Moumtzidou, A. Dimou, and P. King et al. ITI-CERTH participation to TRECVID 2009 HLFE and Search. In *Proc. TRECVID 2009 Workshop*, pages 665–668. 7th TRECVID Workshop, Gaithersburg, USA, November 2009.

[3] A. Moumtzidou, A. Dimou, and N. Gkalelis et al. ITI-CERTH participation to TRECVID 2010. In *Proc. TRECVID 2010 Workshop*. 8th TRECVID Workshop, Gaithersburg, MD, USA, November 2010.

[4] A. Moumtzidou, P. Sidiropoulos, and S. Vrochidis et al. ITI-CERTH participation to TRECVID 2011. In *Proc. TRECVID 2011 Workshop*. 9th TRECVID Workshop, Gaithersburg, MD, USA, December 2011.

[5] A. Moumtzidou, N. Gkalelis, and P. Sidiropoulos et al. ITI-CERTH participation to TRECVID 2012. In *TRECVID 2012 Workshop*, Gaithersburg, MD, USA, 2012.

[6] F. Markatopoulou, A. Moumtzidou, and C. Tzelepis et al. ITI-CERTH participation to TRECVID 2013. In *TRECVID 2013 Workshop*, Gaithersburg, MD, USA, 2013.

[7] N. Gkalelis, F. Markatopoulou, and A. Moumtzidou et al. ITI-CERTH participation to TRECVID 2014. In *TRECVID 2014 Workshop*, Gaithersburg, MD, USA, 2014.

[8] F. Markatopoulou, A. Ioannidou, and C. Tzelepis et al. ITI-CERTH participation to TRECVID 2015. In *TRECVID 2015 Workshop*, Gaithersburg, MD, USA, 2015.

[9] F. Markatopoulou, A. Moumtzidou, and D. Galanopoulos et al. ITI-CERTH participation in TRECVID 2016. In *TRECVID 2016 Workshop*, Gaithersburg, MD, USA, 2016.

[10] F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, and K. Avgerinakis et al. ITI-CERTH participation in TRECVID 2017. In *TRECVID 2017 Workshop*. NIST, USA, 2017.

[11] Konstantinos Avgerinakis, Anastasia Moumtzidou, Damianos Galanopoulos, Georgios Orfanidis, Stelios Andreadis, Foteini Markatopoulou, Elissavet Batziou, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, et al. Iti-certh participation in trecvid 2018. *International Journal of Multimedia Information Retrieval*, 2018.

[12] Konstantinos Gkountakos, Konstantinos Ioannidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. Iti-certh participation in trecvid 2019. In *TRECVID 2019 Workshop*, 2019.

[13] Konstantinos Gkountakos, Damianos Galanopoulos, Marios Mpakratsas, Despoina Touska, Anastasia Moumtzidou, Konstantinos Ioannidis, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. Iti-certh participation in trecvid 2020. In *TRECVID 2020 Workshop*, Gaithersburg, MD, USA, 2020.

[14] Konstantinos Gkountakos, Damianos Galanopoulos, Despoina Touska, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. Iti-certh participation in actev and avs tracks of trecvid 2021. In *TRECVID 2021 Workshop*, Gaithersburg, MD, USA, 2021.

[15] Konstantinos Gkountakos, Damianos Galanopoulos, Despoina Touska, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. Iti-certh participation in actev and avs tracks of trecvid 2022. In *TRECVID 2022 Workshop, Gaithersburg, MD, USA*, 2022.

[16] Damianos Galanopoulos and Vasileios Mezaris. Iti-certh participation in avs task of trecvid 2023. In *TRECVID 2023 Workshop, Gaithersburg, MD, USA*, 2023.

[17] George Awad, Jonathan Fiscus, Afzal Godil, Lukas Diduch, Yvette Graham, and Georges Quénot. Trecvid 2024 - evaluating video search, captioning, and activity recognition. In *Proceedings of TRECVID 2024*. NIST, USA, 2024.

[18] Jiangshan He, Ruizhe Li, Jiahao Guo, Hong Zhang, Mingxi Li, Zhengqian Wu, Zhongyuan Wang, Bo Du, and Chao Liang. WHU-NERCMS AT TRECVID 2023: AD-HOC VEDIO SEARCH (AVS) AND DEEP VIDEO UNDERSTANDING (DVU) TASKS. In *Proceedings of TRECVID 2023*. NIST, USA, 2023.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, , et al. Learning transferable visual models from natural language supervision. In *Proc. of the 38th Int. Conf. on Machine Learning (ICML)*, 2021.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[22] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022.

[23] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[24] J. Xu, T. Mei, et al. MSR-VTT: A large video description dataset for bridging video and language. In *Proc. of IEEE CVPR 2016*, pages 5288–5296, 2016.

[25] Y. Li, Y. Song, L. Cao, J. Tetreault, et al. TGIF: A new dataset and benchmark on animated gif description. In *Proc. of IEEE CVPR 2016*, 2016.

[26] F. Caba Heilbron et al. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proc. of IEEE CVPR 2015*, pages 961–970, 2015.

[27] X. Wang et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proc. of IEEE/CVF ICCV 2019*, pages 4581–4591, 2019.

[28] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt. V3C–a research video collection. In *Proc. of MMM 2019*, pages 349–360. Springer, 2019.

[29] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[30] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

[31] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[32] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1060–1068, 2021.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[34] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, September 2016.

[35] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017.

[36] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking, 2020.

[37] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022.

[38] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.