ITI-CERTH participation in ActEV and AVS Tracks of TRECVID 2021

Konstantinos Gkountakos, Damianos Galanopoulos, Despoina Touska, 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, destousok, kioannid, stefanos, bmezaris, ikom}@iti.gr

Abstract

In this report, the overview of the runs during the TRECVID 2021 by the ITI-CERTH team are presented. ITI-CERTH participated in the Ad-hoc Video Search (AVS) and Activities in Extended Video (ActEV) tasks. For the AVS task, our participation is based on an attention-based cross-modal deep network architecture. As part of training this architecture, we experimented with a new hard-negative mining approach. For the ActEV task, we improve our framework, in terms of more accurate performance, by addressing the classification problem as multi-label rather than a single-label.

1 Introduction

In this work, the work carried out in the context of TRECVID 2021 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 included 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 and in 2019 [12] only in the ActEV task. Lastly, in 2020 [13] ITI-CERTH participated in the AVS, DSDI and ActEV tasks. Taking into account the submissions mentioned above, we aim to evaluate improved algorithms and systems. This year, ITI-CERTH participated in AVS and ActEV tasks. The following sections will present the employed algorithms and the evaluation for the runs we performed in the AVS and ActEV tasks, respectively.

2 Ad-hoc Video Search

The TRECVID 2021 [14] 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. The goal of our participation is twofold. Firstly, we evaluate our attention-based dual encoding network on this year's new ad-hoc textual queries. Secondly, we evaluate a new method for hard-negative mining [15] and compared it with the baseline improved marginal ranking loss.

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

2.1 Approach

We utilize the attention-based dual encoding network presented in [16] as our baseline network. The network is trained to transform an input video-text pair (v, c) into a new joint embedding space. The network consists of two similar sub-networks, one for the video stream and one for the textual one. Each sub-network consists of three levels of encoding, i.e., using mean-pooling, attention-based bi-GRU, and 1d-CNN, layers. Following the state-of-the-art approach [17] [18] [16], the network is trained using the improved marginal ranking loss [18].

As shown in [16] and [15], the combination of multiple models leads to improved performance. In every run, different model configurations are utilized, resulting in different trained models. The models are trained by modifying the following parameters: i) two positions in the architecture are considered for inserting the attention mechanism (textual or visual stream) ii) two textual encodings are used (BERT, W2V+BERT) iii) two optimizers and iv) three learning rates. The resulting 24 models are combined in a late fusion scheme (i.e., averaging a given sample's ranking in the 24 resulting ranking lists).

The second aspect of our study relies on the evaluation of the new hard-negative mining approach. As in [15], we follow an offline-online strategy to exclude potentially-positive samples. At the offline stage, we randomly split the training dataset into batches, similarly to the standard training procedure. In each batch, we compute the cosine similarity score $S_{i,j}^{bert}$, between all possible captions (c_i, c_j) inside the batch. The BERT [19] encoding of the captions is used as the caption representations. Finally, a threshold value p for which $x\%$ of the $S_{i,j}^{bert}$ similarities are higher than p, is computed. At the online stage (training stage), for an video-caption anchor (v_i, c_i), every sample (v_j, c_j) (within the batch) with $S_{i,j}^{bert} > p$ is excluded from the negatives, while every other sample is labeled as negative. Finally, as hard-negative, the negative sample with the highest $S_{i,j}^{bert}$ is selected.

2.2 Submission

The combination of four large-scale video caption datasets: MSR-VTTT [20], TGIF [21], ActivityNet [22] and Vatex [23] is used to train our networks. The V3C1 [24] dataset is utilized to evaluate the networks' performance. The evaluation measure we use is the mean extended inferred average precision (MxinfAP). And finally, as initial frame representations, a ResNet-152 (trained on the ImageNet-11k dataset) is used.

This year we submitted four runs on the AVS 2021 main task and four additional runs for the AVS progress subtask. Overall, we evaluate our methods on 30 different ad-hoc queries. The submitted runs are briefly described below:

- ITI_CERTH.21_run_1: The results of runs 2, 3, and 4 are combined in a late fusion scheme.
- ITI_CERTH.21_run_2: Textual and visual attention-based dual encoding models using multiple textual representations. Models are trained using the improved marginal ranking loss and combined using late fusion.
- ITI_CERTH.21_run_3: Textual and visual attention-based dual encoding models using multiple textual representations. Models are trained using the new hard-negative mining approach when the $x = 1\%$ of the samples are excluded from the hard-negative mining because of being treated as potentially positive samples.
- ITI_CERTH.21_run_4: Similar to run 3 with $x = 2\%$.

2.3 Experimental Results

Table 1 summarizes the evaluation results of our runs for the main AVS task as well as the progress subtask. From the presented results, we can see that that our new hard-negative mining approach is meaningful, in the sense that it doesn't negatively affect the performance, although by itself it doesn't improve it either. When models that have been trained using the improved marginal ranking loss and the new hard-negative mining approach are combined, even in a simple late fusion scheme as in ITI_CERTH.21_run_1, we observe a small increase in performance, in the main AVS task. The sub-optimal design choice in this run is the simple late fusion scheme: as shown in [15], using the new

hybrid model combination strategy proposed there instead of simple late fusion would have resulted in a much more pronounced increase in performance.

Table 1: Mean Extended Inferred Average Precision (MXinfAP) for all submitted runs for the fully-automatic AVS task.

AVS task:	Main	Progress
ITI_CERTH.21 run_1	**0.232**	0.225
ITI_CERTH.21 run_2	0.227	0.226
ITI_CERTH.21 run_3	0.225	0.226
ITI_CERTH.21 run_4	0.227	0.226

3 Activities in Extended Video

Activity recognition is the task of analysing multimedia resources like videos and photos and identifying the specific movement or actions of a person and other objects. A widespread application refers to security systems, including surveillance cameras in indoor and outdoor environments. Activity recognition in those systems deals with plenty of challenges inserted by the visual footage's untrimmed nature, like the large field of view, the multiple involved activities, the varying length of co-occurring activities, and the multiple objects involved within each activity. In conjunction with the need for a real-time response, the latter makes difficult the processing and analysis by humans. Hence, fully-automated methods to recognise and localise spatially and temporally activities in extended untrimmed videos are on demand. In this direction, the Activities in Extended Videos challenge (ActEV) encourage the research of real-time automatic activity detection methods in surveillance scenarios. Thus, an extensive collection of untrimmed surveillance videos such as VIRAT [25] dataset jointly with an evaluation plan are provided, making possible the standardised evaluation of the related methods under a unified framework.

In this work, the proposed approach consists of a three-step pipeline: object detection, tracking, and activity classification. The method utilises YOLOv4 [26] architecture for object detection and Euclidean distance for tracking, resulting in the creation of spatio-temporal tubelets of the detected objects and thereby of the proposed activities, while the activity recognition employs 3D-ResNet [27] assigning possible labels to each proposed activity. The rest of the section is organised as follows: The following subsection gives an overview of the steps comprised in the method, outlining the parts of the object detection and activity recognition, correspondingly. The section concludes with a report about the submitted systems and a discussion of the results.

3.1 Approach

The task of activity recognition and localization in extended videos has the following formulation. Given a set of videos $V = \{v_i\}$, a method should be able to output a set of activities $A = \{a_i\}$ for all the videos in the set. Each one of the activities a_i is described by a type t_i, according to a set of classes provided by the corresponding annotation files and are depicted in Table 2, and a spatio-temporal area l_i within the video in which it occurs.

Considering the challenges on detected activities in surveillance videos, this work focuses on effectively identifying the activities co-occurring simultaneously and conducted by different actors-objects, as well as the concurrent activities performed by the same object. Thus, an object detector in conjunction with a tracker is employed to generate spatio-temporal tubelets for every detected object. The classification of the proposed spatio-temporal tubelets follows the creation of the Extended Activity Bounding Box (EABBox) [13] for every tubelet considering the spatial displacement of an object during its action. In particular, the extended tubelets can be described as a tuple $p_i = (x_{left}, y_{top}, width, height, t_{start}, t_{end})$ consisting of the spatial and temporal boundaries of an object. Similar to our previous work [13], the process of EABBox creation aims to capture the whole

field of view of the object's movement providing useful information for the final classification. Further details of the pipeline are described in the following sections.

Table 2: Activity classes in ActEV challenge 2020.

Activity Classes		
person_closes_facility _or_vehicle_door	person_closes_trunk	vehicle_drops_off_person
person_enters_facility_or_vehicle	person_exits_facility_or_vehicle	person_interacts_object
person_loads_vehicle	person_opens_trunk	person_opens_facility _or_vehicle_door
person_person_interaction	person_pickups_object	vehicle_picks_up_person
person_pulls_object	person_pushs_object	person_rides_bicycle
person_sets_down_object	person_talks_to_person	person_carries_heavy_object
person_unloads_vehicle	person_carries_object	person_crouches
person_gestures	person_runs	person_sits
person_stands	person_walks	person_talks_on_phone
person_texts_on_phone	person_uses_tool	vehicle_moves
vehicle_starts	vehicle_stops	vehicle_turns_left
vehicle_turns_right	vehicle_makes_u_turn	

3.1.1 Object Detection

In this section, the spatio-temporal tubelets generation is introduced, elaborating the videos in a frame-wise manner. This stage includes the detection and tracking methods to identify the candidate objects within each video and propose the spatio-temporal tubelets for activity classification. Taking into account the fast and accurate performance of YOLOv4 [26] in real-time applications, achieving 43.5% Average Precision (AP) for the Microsoft COCO [28] dataset at a real-time speed of approximately 65 FPS on Tesla V100; we adopt it as means to capture the spatial boundaries of the objects in every frame of a video. YOLOv4 [26] composes an improved version of YOLOv3 [29] as it is typified as a state-of-the-art detector combining a fast operating speed in production systems and optimisations for parallel computations. For initialisation purposes, we utilised the pre-trained model of YOLOv4 [26] using Microsoft COCO [28] which consists of objects classes relevant to the objects that participate in the activities of the challenge's dataset, including but not limited to "person", "car" and "truck". An enhanced configuration of the framework was introduced including the fine-tuning of YOLOv4 [26] using the VIRAT [25] dataset. The latter was split into training and validation sets according to the annotations of ActEV challenge. We fine-tuned the model for 20 epochs to target only the detection of vehicle and person objects. Hence, the time-consuming post-processing steps were discarded as the generated predictions are more accurate, considering only the desired objects' types. This methodology is partially adopted in our relevant research described in [30], where there is a detailed qualitative and quantitive explanation of the improvements that fine-tuning provides. The detected objects for each frame are described by a bounding box and the corresponding confidence score.

Along with object detection, an object tracker was adopted in order to link object detections over time and calculate their trajectories. For this purpose, the Euclidean distance was deployed as a metric to measure the distance among the detected objects in subsequent frames. The algorithm calculates the distance between the current object on frame t and the objects on frame $t+1$. An adjacent object of frame $t+1$ is selected as the future position of the current object when its distance is above a threshold value. The Euclidean distance is calculated between the centroids of the objects' bounding boxes. The result of this stage is the spatio-temporal tubelets generation for each detected object.

3.1.2 Post-processing

In order to have ready the proposed spatio-temporal tubelets for the final step of activity classification, an additional procedure is required, that of EABBox creation. The latter elaborates the spatial

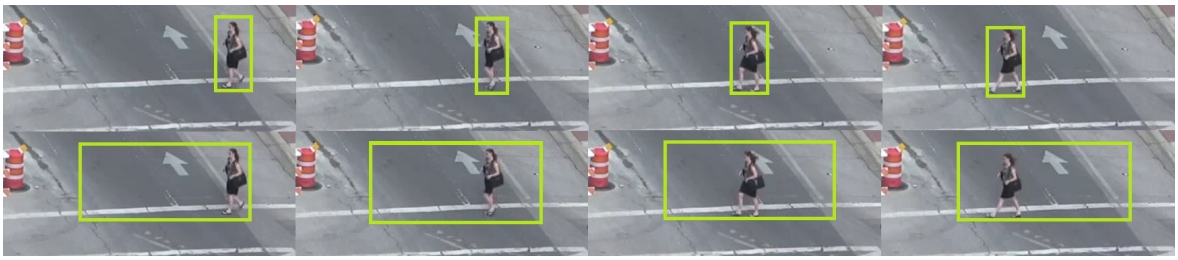

Figure 1: First row: frames sequence with the detected object by YOLOv4 [26]. Second row: frames sequence after EABBox creation.

boundaries of each spatio-temporal tubelet in order to extend the field of view of the object considering its whole trajectory. Thus, the union of the separated bounding boxes of each tubelet is calculated as it is occurred from the spatial displacement of the object during its movement; this is illustrated in figure 1. Some benefits of EABBox creation, are the minimisation of the effect that the cropping has on the frames reshaping them according to the size of a bounding box, which frequently leads to a stretched and deform illustration of the objects. Another reason is that useful information from the background is included in the extended version of bounding boxes which can be helpful in the following procedure of activity classification. Finally as already mentioned, a spatio-temporal tubelet proposal is defined by a tuple $p_i = (x_{left}, y_{top}, width, height, t_{start}, t_{end})$ consisting of its spatial and temporal boundaries.

3.1.3 Activity Recognition

The final stage of the process is the activity classification. Given the spatio-temporal tubelets of the tracked objects, a 3D-ResNet [27] was employed to label the proposed tubeletes. The selection of this model was highly correlated with its ability to process the data in a 3D state also exploiting the temporal information that they include in contrast to 2D Convolutional Neural Networks (CNNs) that can learn only spatial correlations. The model was trained and validated using the provided videos of the VIRAT [25] dataset, according to the annotations of ActEV for splitting it into training and validation sets, respectively. We follow the approach utilised in [30], classified the proposed tubeletes in a multi-label manner. In other words, the employed classifier assigns to each proposed spatio-temporal tubelet more than one label to their whole trajectory or different temporal sub-parts of their trajectory. A weighted binary cross-entropy loss function was incorporated regarding the model's architecture to transform the problem into a multi-label objective and deal with the unbalanced dataset. Lastly, a Soft-NMS [31] is applied to refine the classified activities proposals.

3D-ResNet is a deep neural network model which comprises a 3D-convolutional-based architecture and achieves fast processing for activity recognition in a real-time state, running in batch-of-frames. The architecture consists of four sequential bottleneck blocks, where each block is composed of three 3D-convolution layers (with variant kernel sizes), batch normalisation, and ReLU activation layers. The temporal dimension of the input is set equal to 16. For the model's initialisation, the weights of the Kinetics dataset [32] were pre-loaded as it covers a large number of human activity classes.

3.2 Submission

In this section, we present the two systems that were submitted and the evaluation results of the challenge in figure 2 and on table 3.

- M4D_2021-baseline: This system uses YOLOv4 [26], fine-tuned in the training and validation set of ActEV, for object detection and the Euclidean distance for tracking. Subsequently, the corresponding data sets were used for training and validating the activity classifier 3D-Resnet [27] to recognise the activities of the tracked objects, performing in a multi-label manner.
- M4D_2021-M4D_2021_S1: This system combines the first system with some additional steps of improvements in the output set of activities proposals. The latter is conducted using Soft-NMS [31] to refine the temporal boundaries of the recognised activities.

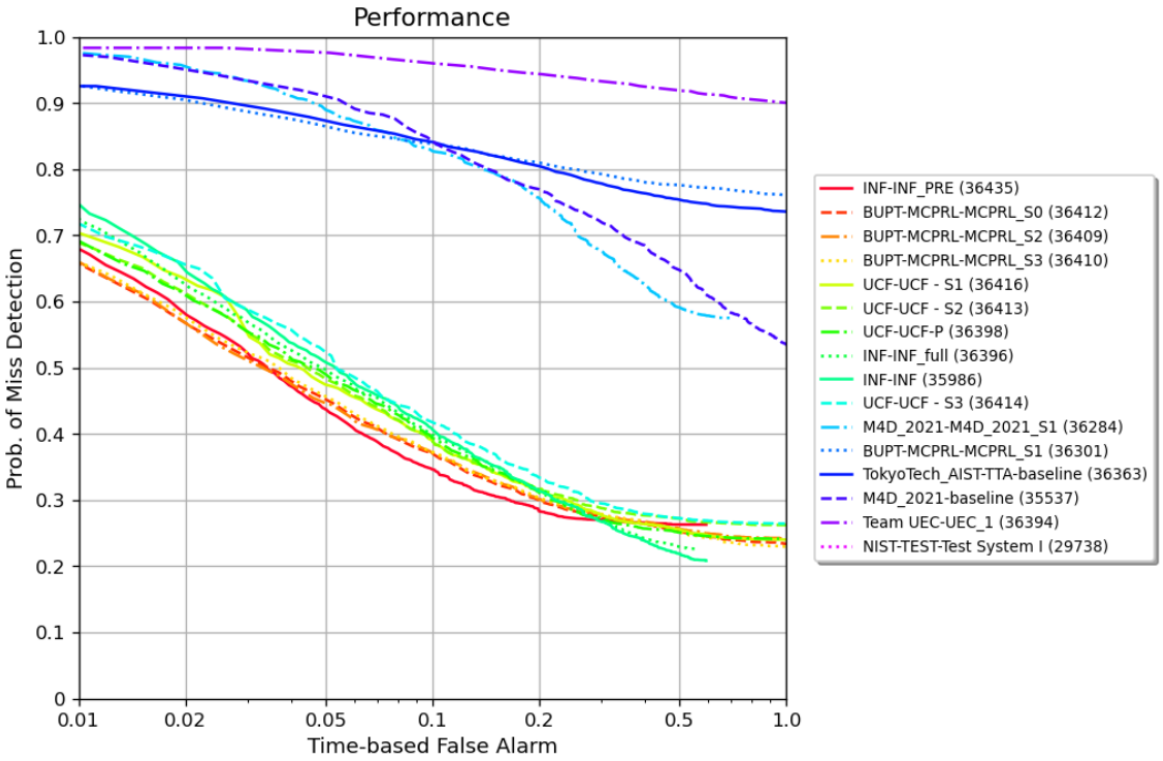

Figure 2: ActEV 2021 Leaderboard.

Table 3: Our ActEV challenge 2021 results, ranked using PARTIAL AUDC as primary metric.

System	nAUDC@0.2T_{fa}	Mean-P_{miss}@0.15T_{fa}	Mean-W_P_{miss}@ 0.15R_{fa}
M4D_2021-baseline	0.85484	0.79732	0.87719
M4D_2021-M4D_2021_S1	0.84658	0.79410	0.88521

3.3 Experimental Results

In this section, further discussion about the performance of our both submitted systems is reported. The significant change that discriminates our submissions in 2021 from those in 2020 is the different manner of training and applying activity classifiers, namely the transition from multi-class to multi-label approach. In particular, compared to our reported results [13] on ActEV 2020, we succeed to decrease the primary metric, PARTIAL AUDC, in both of our submissions on ActEV 2021. This is due to the fact that a multi-label approach of training corresponds more efficiently to the process of assigning labels to the proposed spatio-temporal tubelets, as each one of them is related to an object which can perform more than one activity at the same time or at different time intervals during its trajectory. Along with the multi-label activity classification, YOLOv4 [26] fine-tuning influences in a positive direction the results, as it outputs only the objects types (person and vehicle) that participate in the annotated activities, excluding the time-consuming post-processing steps for the objects' detections refinement. Lastly, regarding our second submission, M4D_2021-M4D_2021_S1, we observed that the addition of the Soft-NMS [31] algorithm improves the results as it offers the possibility to eliminate duplicate activities which affect negatively the results.

4 Conclusions

In this paper, the evaluation of ITI-CERTH during the TRECVID 2021 challenge [14] is reported. ITI-CERTH participated in the AVS and ActEV tasks to evaluate new techniques and algorithms. Regarding the AVS task, we utilized an attention-based cross-modal network to learn a new joint feature space for the text and video instances. We experimented with a new hard-negative mining approach for training this network architecture; and, we contemplated on possible approaches to combining the results of multiple trained instances of this network architecture that were trained using different settings for the hard-negative mining and for other training parameters. At the ActEV task, a method based on an object detector, an object tracker, and a multi-label activity classifier is presented. The method relies on a real-time object detector and a 3D-CNN activity classifier. Though the results are not expected, some aspects of the process seem promising. We plan to intensify our effort for improved systems and proper model training in the future.

5 Acknowledgements

This work was partially supported by the European Commission under contracts H2020-786731 CONNEXIONs, H2020-833115 PREVISION, and H2020-832921 MIRROR.

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] George Awad, Asad A. Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas Diduch, Jeffrey Liu, Yvette Graham, Gareth J. F. Jones, , and Georges Quénot. Evaluating multiple video understanding and retrieval tasks at trecvid 2021. In *Proceedings of TRECVID 2021*. NIST, USA, 2021.

[15] D. Galanopoulos and V. Mezaris. Hard-negatives or Non-negatives? A hard-negative selection strategy for cross-modal retrieval using the improved marginal ranking loss. In *2021 IEEE/CVF ICCVW*, 2021.

[16] D. Galanopoulos and V. Mezaris. Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In *Proc. of the ACM Int. Conf. on Multimedia Retrieval*, (ICMR '20). ACM, 2020.

[17] J. Dong, X. Li, C. Xu, S. Ji, Y. He, et al. Dual encoding for zero-example video retrieval. In *Proceedings of IEEE Conf. CVPR 2019*, pages 9346–9355, 2019.

[18] F. Faghri, D. J. Fleet, et al. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proc. of the British Machine Vision Conference (BMVC)*, 2018.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018.

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

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

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

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

[24] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. V3C–a research video collection. In *International Conference on Multimedia Modeling*, pages 349–360. Springer, 2019.

[25] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011.

[26] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[27] 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.

[28] 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.

[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018.

[30] Konstantinos Gkountakos, Despoina Touska, Konstantinos Ioannidis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. Spatio-temporal activity detection and recognition in untrimmed surveillance videos. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 451–455, 2021.

[31] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[32] 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.