# VIREO @ TRECVid 2020 Ad-hoc Video Search

Jiaxin Wu, Phuong Anh Nguyen, Chong-Wah Ngo

*Department of Computer Science, City University of Hong Kong*
{jiaxin.wu, panguyen2-c}@my.cityu.edu.hk,
cscwngo@cityu.edu.hk

## Abstract

In this paper, we summarize our submitted runs and results for Ad-hoc Video Search (AVS) task at TRECVid 2020 [1].

**Ad-hoc Video Search (AVS):** We applied two video search systems for AVS: a dual-task video search system [2] and a concept-based video search system [3]. The dual-task model [2] learns feature embedding and concept decoding simultaneously in an end-to-end training manner. In contrast, the concept-based video search system [3] is trained on off-the-shelf concept banks. Our aims is to analyze the advantages and shortcomings of these video search approaches. We submitted four automatic runs and four manual runs for both main task and progress subtask. Besides, we also submitted one novelty run for the main task. We briefly summarize our runs as follows:

- *F_D_C_D_VIdeoREtrievalGrOup.20_1*: This automatic run attains the mean xinfAP= 0.206 on the main task and xinfAP= 0.230 on the progress subtask. This run is based on our recently proposed feature embedding technique using the dual-task model. The whole query is directly input to the dual-talk model to get the textual embedding. The search result is output based on the cosine similarity of the textual embedding and all video embeddings.
- *F_D_C_D_VIdeoREtrievalGrOup.20_2*: This automatic run attains the mean xinfAP= 0.183 on the main task and xinfAP= 0.147 on the progress subtask. This run is based on the concept decoding using the dual-task model. The query is mapped into query tokens, and the tokens are used as indexes to find matches in the decoding concept lists of videos.
- *F_D_C_D_VIdeoREtrievalGrOup.20_3*: This automatic run obtains the mean xinfAP= 0.229 on the main task and xinfAP= 0.248 on the progress subtask. This is the best automatic run of our submissions. This run is based on the late fusion of feature embedding and concept decoding using the dual-task model.
- *F_D_C_D_VIdeoREtrievalGrOup.20_4*: This automatic run attains the mean xinfAP= 0.113 on the main task and xinfAP= 0.134 on the progress subtask. This run is based on our previous concept-based video search system [3] which is trained on multiple off-the-shelf classification datasets.
- *M_D_C_D_VIdeoREtrievalGrOup.20_1*: This manual run applies the same system with the same settings presented in the run *F_D_C_D_VIdeoREtrievalGrOup.20_1*. The difference is that the user parses and categorizes the query manually at the beginning of the process. This human intervention degrades the performance of the main task from 0.206 to 0.203 but improves the performance from 0.230 to 0.266 for the progress subtask.

- *M_D_C_D_VIdeoREtrievalGrOup.20_2*: This manual run is based on the same system with the same settings presented in the run *F_D_C_D_VIdeoREtrievalGrOup.20_2*. Starting from the list of automatically selected query tokens for each query, the user screens the concept list and manually modifies the query tokens. Consequently, the performance degrades from 0.187 to 0.177 for the main task but the result rises significantly from 0.147 to 0.230 for the progress subtask.
- *M_D_C_D_VIdeoREtrievalGrOup.20_3*: This manual run is the fusion of the previous two manual runs. In this run, the performance is improved by 0.04 xinfAP for the main task. The result boosts from 0.248 to 0.288 for the progress subtask.
- *M_D_C_D_VIdeoREtrievalGrOup.20_4*: This manual run uses the same system with the same settings presented in the run *F_D_C_D_VIdeoREtrievalGrOup.20_4*. Starting from the list of automatically selected concepts for each query, the user screens the concept list and removes unrelated or unspecific concepts to refine the result. This step helps improving the mean xinfAP significantly from 0.113 to 0.223 for the main task. The result is also improved from 0.134 to 0.207 for the progress subtask.
- *M_D_N_D_VIdeoREtrievalGrOup.20_5*: This manual run uses the same system and the same manual queries with the run *M_D_C_D_VIdeoREtrievalGrOup.20_2*. The difference is that we manually added some concepts in each query to prune the result returned by the embedding search. This run attains mean xinfAP= 0.138 for the main task.

# 1 Ad-hoc Video Search (AVS)

Concept-based search has been the mainstream approaches for ad-hoc video search (AVS) task since the benchmarking [4, 5, 6]. It relies on concept detectors to detect several kinds of concepts such as person, object, action, place in the videos, and then use the detected concepts as indexes of videos to retrieve. As concept characterizes high-level semantics, the search process is explainable and predictable. However, the process of concept-based search is hindered by the challenging issues, e.g., concept selection and concept fusion [7, 8]. In contrast, embedding search bypasses these issues and measures the similarity of two modalities (query text and video) in a joint space. It has shown its powerful retrieval ability on the AVS and has become the mainstream approach since year 2018 [9, 10, 11]. However, as embedding search is performed in the feature space, it is hard to interpret, and the result is not predictable.

As these two kinds of approaches have their own merits and shortcomings, we have studies a new network [2] which combines embedding search and concept-based search in a unified architecture. Figure 1 displays an overview of the network. It includes a stream for embedding feature learning between two modalities. Besides, this network also involves another stream for concept decoding. Consequently, each video is indexed with an embedding feature and a list of concepts for search, and the concept list can be seen as an interpretation for the embedding feature.

The dual-task model provides three schemes for retrieval: embedding search, concept search, and fused search. We have used these three schemes for this year's AVS benchmarking which are VIdeoREtrievalGrOup.20_1&2&3. To compare with the traditional concept-based models, we also include our previous concept-based search system [3] as the fourth run (VIdeoREtrievalGrOup.20_4). We have utilized three video caption datasets: TGIF [12], MSR-VTT [13] and VidOR-MPVC, to train the dual-task model and the remaining settings are the same as [2]. The size of vocabulary in the dual-task concept bank is 11,613.

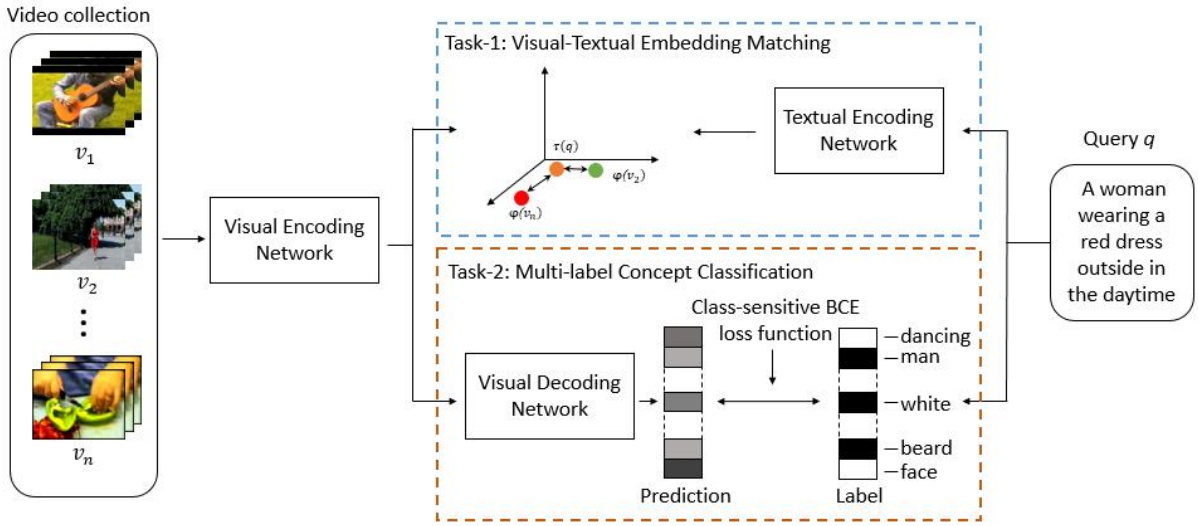


Figure 1: The architecture of the dual-task model. There are two channels: one is for embedding feature learning, and the other is for concept decoding. Both channels are trained in an end-to-end manner to achieve consistency in the visual embedding.

## 1.1 Detail descriptions

### 1.1.1 F_D_C_D_VIdeoREtrievalGrOup.20_1

This run utilizes the dual-task embedding search model. The trained textual encoding network projects the user query to the same latent space as the video embeddings. Then, the similarity of this query and all videos are measured in the common space. A score is computed for each video based on the cosine similarity:

$$score_{embedding}(q, v_i) = sim(\tau(q), \phi(v_i)) \tag{1}$$

where $\tau(q)$ and $\phi(v_i)$ denote the embeddings of the query and a video in the joint space, respectively.

### 1.1.2 F_D_C_D_VIdeoREtrievalGrOup.20_2

In this run, we utilize the dual-task concept system to retrieve videos. Firstly, the input query is mapped into query tokens based on the dual-task concept bank. Then, these tokens are used as keys to find matches on the decoding concept lists of video segments. A concept score will be computed for each video:

$$score_{concept}(q, v_i) = sim(c_q, \hat{y}(v_i)), c_q \in \{0, 1\}^n, \hat{y}(v_i) \in \mathbb{R}^{n+} \tag{2}$$

where $\hat{y}(v_i)$ is the predicted probability of decoding concepts of the video $v_i$. The notation $n$ is the size of dual-task concept bank. A positive value in the query vector $c_q$ means that concept is selected as index to search. All videos are sorted based on their concept scores.

### 1.1.3 F_D_C_D_VIdeoREtrievalGrOup.20_3

We apply the dual-task fused search to submit the result in this run. The fused search is achieved by lately fusing the embedding and concept searches by a linear function as:

$$score_{fused}(q, v_i) = \theta * score_{concept}(q, v_i) + (1 - \theta) * score_{embedding}(q, v_i). \tag{3}$$

We use $\theta = 0.3$ in this evaluation. This run contributes the best performance of our runs this year on both main task and progress subtask.

#### 1.1.4 F_D_C_D_VIdeoREtrievalGrOup.20_4

In this run, we have re-run our previous concept-based system [3]. The system is trained on multiple concept datasets includes ImageNet12988 [14], ImageNet1000 [15], RC497 [16], Kinetics60 [17], OpenImageV4 [18], SIN346 [19], place365 [20]. The concept bank size is 16,396.

## 2 Results analysis

In this year's AVS benchmarking, the evaluation is conducted on the V3C1 dataset [21] and two groups of queries. The first group contains twenty new main queries released this year. The other group includes ten progress queries released last year. The use of progress queries is to facilitate the performance comparison between the submissions in the years 2019 and 2020.

Figure 2 shows the mean extended inferred average precision (xinfAP) of 20 new queries for 26 submissions this year, and our results are highlighted in red. Overall, we rank third among nine teams. Our model suffers from out-of-vocabulary and out-of-training-examples problems. For example, query-644 *Find shots of sailboats in the water* and query-653 *Find shots of group of people clapping*. As there are few training samples of "sailboats" and "people clapping", our dual-task model only manages to achieve 0.001 and 0.061 xinfAPs on these two queries. Our fused search gets better results than our embedding search and concept search, as those correct video segments which are agreed by both embedding and concept searches will have higher combined scores, and they will be lifted up to the top of the search list. For example, for the query-658 *Find shots of two or more people under a tree*, embedding and concept searches find 38 and 47 positive video segments respectively in their top-100 list, and fused search manage to have 51 correctness in the top 100. It means some positive video segments out of top-100 are brought forward, leading to a higher xinfAP. Moreover, our dual-task concept search exceeds our previous concept model by 0.07 mean xinfAP for the main queries, although the size of our concept bank in the dual-task model is smaller than our previous concept model (11,613 versus 16,396). The good performance of

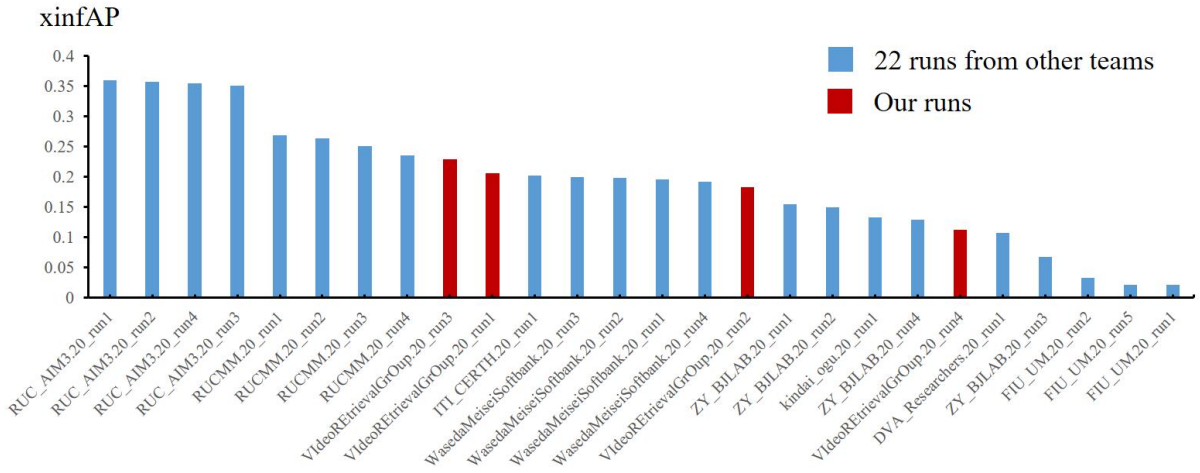


Figure 2: Performance comparison of our submissions and other teams in fully-automatic runs on main queries.

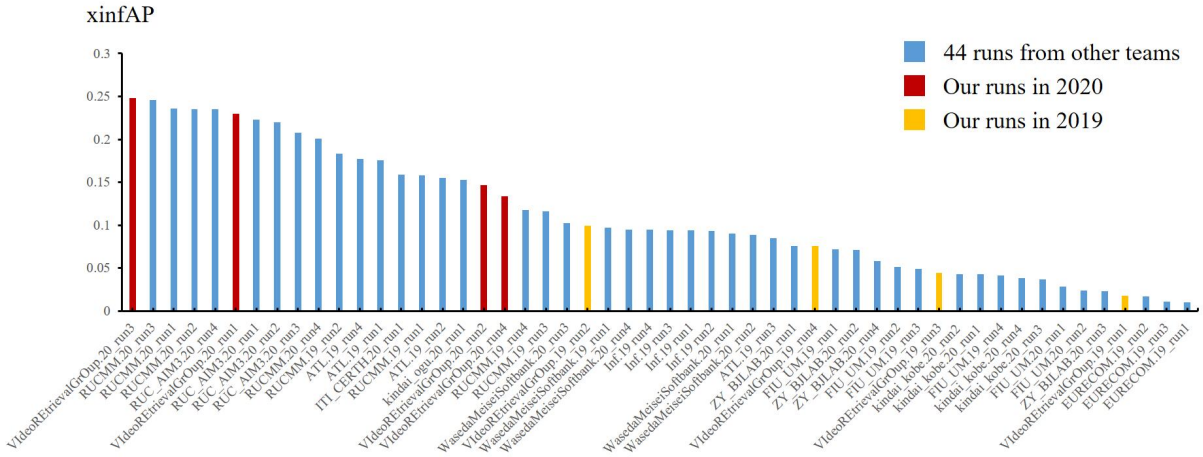


Figure 3: Performance comparison of our submissions and other teams in fully-automatic runs on progress queries.

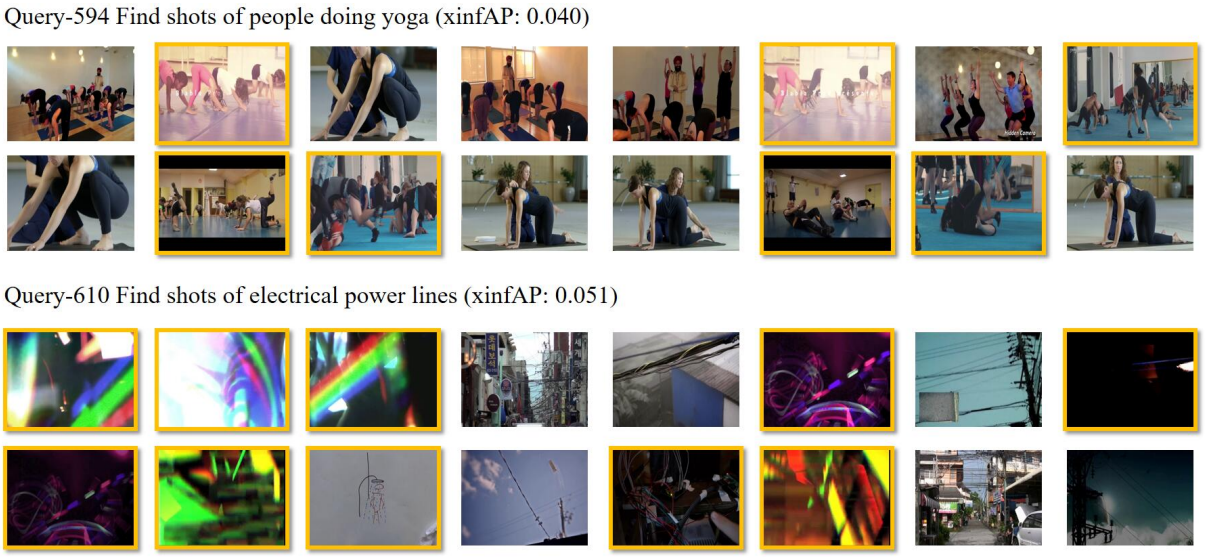


Figure 4: The top-16 retrieved results for two progress queries by the dual-task fused model. The false positives are highlighted in yellow.

our current concept model is evidenced on some queries such as query-648 *Find shots of a man in blue jeans outdoors* and query-660 *Find shots of train tracks during the daytime*. Our new concept model achieves twofold increases on these two queries because of higher accuracy on the concepts, e.g., "blue jeans" and "train tracks". However, on some queries such as query-644 and query-653, our previous concept system performs better as the off-the-shelf datasets have many training samples on "sailboats" and "people clapping". It is worth mention that, the performances of our previous concept model are significantly improved in manual runs. In contrast, our recently proposed dual-task embedding model and concept model do not get benefit on some manually edited queries. The main reason is that we use different strategies on two systems in modifying queries. As our previous concept system suffers a lot from out-of-word problem, our strategy is to find concept synonyms to get rid of this problem. Meanwhile, as

this system starts from finding nothing to find somethings, the improvement could always be seen. On the other hand, for the new model, we not only try rephrasing the queries but also try adding constrains to the queries. We find that our dual-task model will be benefit from rephrasing and be worse when having more restrictions on the queries. For example, for the query-660 *Find shots of train tracks during the daytime*, we manually modified the query by adding the concept "outdoors". The performance of the dual-task concept model drops from 0.466 to 0.159. The big degradation could also be found when we added "interior room" to the query-659 *Find shots of a church from the inside*. Besides, as the results of embedding model are unpredictable, the slight drop is also seen in the dual-task embedding model for the manual run. For example, the performance drops from 0.087 to 0.036 when changing the query *Find shots of one or more people drinking wine* to *Find shots of people drinking wine*, although two queries have the same meaning.

Figure 3 shows the results of the 10 progress queries evaluated this year. Our submissions this year are highlighted in red, and the submissions last year are marked in yellow. It is obvious that our current system is better than the previous system. The improvement is significant in our dual-task fused search (VIdeoREtrievalGrOup.20_run3). Its performance has doubled the best result of our previous concept-based system (VIdeoREtrievalGrOup.19_run2). Our fused model works well on those complex queries such as query-593 *Find shots of one or more women models on a catwalk demonstrating clothes* and query-602 *Find shots of a person jumping with a bicycle*. One of the main reasons may be due to the biGRU [22] in the dual-task textual encoder which could encode the sequence information of query. However, it has failed terribly in three queries: query-594 *Find shots of people doing yoga*, query-606 *Find shots of people queuing* and query-610 *Find shots showing electrical power lines*. It only manages to gain 0.040, 0.002, and 0.051 xinfAP scores on them. For query-606, the reason is lacking training samples in "queuing". For query-594, the search results are contaminated with videos of similar actions, such as sketching and wrestling as shown in Figure 4. The false positives (highlighted in yellow) could be effectively pruned if we search for the object "yoga mat" instead of the action "doing yoga". The performance will be improved significantly from 0.040 to 0.232. For query-610, the model seems confusing of what is "electrical power lines". Sometimes, it finds "electric light lines" or "electrical wires" as results as shown in Figure 4. For this cases, more positive examples of "electrical power lines" are needed to let the model understand this concept clearly.

## 3 Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61872256).

## References

[1] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, and G. Quénot, "Trecvid 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains," in *Proceedings of TRECVID 2020*.

[2] J. Wu and C.-W. Ngo, "Interpretable embedding for ad-hoc video search," in *Proceedings of the ACM Conference on Multimedia*, 2020.

[3] P. A. Nguyen, J. Wu, C.-W. Ngo, F. Danny, and H. Benoit, "Vireo-eurecom @ trecvid 2019: Ad-hoc video search," in *Proceedings of the TRECVID 2019 Workshop*, 2019.

[4] V.-T. Nguyen, D.-D. Le, B. Renoust, T. D. Ngo, M.-T. Tran, D. A. Duong, and S. Satoh, "NII-HITACHI-UIT at TRECVID 2016 ad-hoc video search: Enriching semantic features using multiple neural networks," in *Proceedings of the TRECVID 2016 Workshop*, 2016.

[5] F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, T. Mironidis, V. Kaltsa, A. Ioannidou, S. Symeonidis, K. Avgerinakis, S. Andreadis, I. Gialampoukidis, Stefanos, Vrochidis, A. Briassouli, V. Mezaris, I. Kompatsiaris, Ioannis, and Patravs, "Iti-certh participation in trecvid 2016," in *Proceedings of the TRECVID 2016 Workshop*, 2016.

[6] J. Liang, P. Huang, L. Jiang, Z. Lan, J. Chen, and A. Hauptmann, "Informedia @ trecvid 2016 med and avs," in *Proceedings of the TRECVID 2016 Workshop*, 2016.

[7] K. Shirahama, D. Sakurai, T. Matsubara, and K. Uehara, "Kindai university and kobe university at trecvid 2019 avs task," in *Proceedings of the TRECVID 2019 Workshop*, 2019.

[8] K. Ueki, K. Hirakawa, K. Kikuchi, T. Ogawa, and T. Kobayashi, "Waseda meisei at trecvid 2017:ad-hoc video search," in *Proceedings of the TRECVID 2017 Workshop*, 2017.

[9] X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong, "W2VV++: Fully deep learning for ad-hoc video search," in *Proceedings of the ACM International Conference on Multimedia*, 2019.

[10] X. Li, J. Ye, C. Xu, S. Yun, L. Zhang, X. Wang, R. Qian, and J. Dong, "Renmin university of china and zhejiang gongshang university at trecvid 2019: Learn to search and describe videos," in *Proceedings of the TRECVID 2019 Workshop*, 2019.

[11] X. Wu, D. Chen, Y. He, H. Xue, M. Song, and F. Mao, "Hybrid sequence encoder for text based video retrieval," in *Proceedings of the TRECVID 2019 Workshop*, 2019.

[12] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo, "Tgif: A new dataset and benchmark on animated gif description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[13] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[14] P. Mettes, D. C. Koelma, and C. G. Snoek, "The imagenet shuffle: Reorganized pre-training for video event detection," in *Proceedings of the ACM on International Conference on Multimedia Retrieval*, 2016, pp. 175–182.

[15] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[16] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel, "Creating havic: Heterogeneous audio visual internet collection," in *Proceedings of the Eight International Conference on Language Resources and Evaluation*, 2012.

[17] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *ArXiv*, vol. abs/1808.01340, 2018.

[18] A. Kuznetsova, H. Rom, N. Alldrin, J. R. R. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, "The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale," *CoRR*, vol. abs/1811.00982, 2018.

[19] G. Awad, C. G. M. Snoek, A. F. Smeaton, and G. Quénot, "TRECVid Semantic Indexing of Video: A 6-Year Retrospective," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 2016, p. 22, 2016.

[20] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proceedings of the International Conference on Neural Information*, 2014, pp. 487–495.

[21] F. Berns, L. Rossetto, K. Schoeffmann, C. Beecks, and G. Awad, "V3c1 dataset: An evaluation of content characteristics," in *Proceedings of the International Conference on Multimedia Retrieval*, 2019, pp. 334–338.

[22] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.