# A Multi-Stream Approach for Video Understanding

Lutharsanen Kunam
University of Zurich
Zurich, Switzerland
lutharsanen.kunam@uzh.ch

Luca Rossetto
University of Zurich
Zurich, Switzerland
rossetto@ifi.uzh.ch

Abraham Bernstein
University of Zurich
Zurich, Switzerland
bernstein@ifi.uzh.ch

## ABSTRACT

The automatic annotation of higher-level semantic information in long-form video content is still a challenging task. The Deep Video Understanding (DVU) Challenge aims at catalyzing progress in this area by offering common data and tasks. In this paper, we present our contribution to the 3rd DVU challenge. Our approach consists of multiple information streams extracted from both the visual and the audio modality. The streams can build on information generated by previous streams to increase their semantic descriptiveness. Finally, the output of all streams can be aggregated in order to produce a graph representation of the input movie to represent the semantic relationships between the relevant characters.

## CCS CONCEPTS

• **Information systems** → *Evaluation of retrieval results*; Video search; • **Computing methodologies** → Video summarization.

## KEYWORDS

Video Understanding, Multimodal Analysis

**ACM Reference Format:**
Lutharsanen Kunam, Luca Rossetto, and Abraham Bernstein. 2022. A Multi-Stream Approach for Video Understanding. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3503161.3551567

## 1 INTRODUCTION

While the means for the extraction of semantic information from various forms of multimedia content has made rapid progress in recent years, resulting in a multitude of methods capable of annotating isolated instances of semantic concepts, understanding of long-form content remains a challenging task. The purpose of the Deep Video Understanding Challenge is to offer a forum for the advancement in this area. The goal of the challenge is to transform long-form video into a graph representation that contains all semantically relevant concepts, such as characters and locations and accurately captures their relations as depicted in the film. In this paper, we describe our participation to the third edition of this challenge. Our pipeline tasked with the generation of this graph uses an iterative multi-stream approach, building on the output of previous streams to increasingly generate more abstract semantic information.

The remainder of this paper is structured as follows: Section 2 provides a brief overview of approaches used in previous editions of the challenge. Section 3 describes the pipeline used in our challenge participation and provides some details on how the graphs are constructed and the queries are processed. Section 4 then shows some insights gained from the output of the pipeline and Section 5 discusses some identified limitations and open challenges. Finally, Section 6 concludes.

## 2 RELATED WORK

This section provides an overview of the approaches used in the previous two iterations of the challenge from 2020 and 2021.

One of the methods introduced in 2020 in [2] is based on a multimodal approach, using a range of different information channels for the prediction of semantic relationships between entities. These channels include local and global visual feature descriptors, kinetic action descriptors, as well as tonal and semantic features of the spoken dialogue. The method first segments the input movie into shots which are then aggregated into semantically coherent scenes. It then maps the conversational discourses to interacting co-located objects and fuses them with kinetic action embeddings in each scene. From this, it generates a combined probability distribution representation for interacting entities which is used to predict their semantic relationship.

The approach in [3] also uses shot segmentation to generate a series of information units to be analyzed individually. For each shot, characters are detected using face identification and locations and concepts are detected using a global visual descriptor. Every pair of entities detected in a scene is assumed to have a relationship of some sort. These individual relationships are then identified using a series of hierarchical binary classifiers trained on a combination of features extracted from the automatically transcribed dialogue.

The best performing method of 2020 [24] segments the input video into semantic scenes directly and uses local visual descriptors for location recognition and face and body tracking for person identification in every scene. A combination of visual features of the characters, and semantic and tonal speech features is then used to predict the relations.

The approach introduced in [25] in 2021 uses a feature comprised of six components for the classification of relations. For every scene, it uses audio features and embeddings of the transcribed text as well as embeddings of the space-time volumes of the entire scene, the volumes occupied by the subject, and the object of the relation, and the embedding of the bounding volume of these volumes.

The best performing approach of 2021 [1] departs from the commonalities of the previously discussed ones to some degree by


*MM '22, October 10–14, 2022, Lisboa, Portugal*

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3551567

taking inspiration from the natural language processing community. They describe their method as *"a zero-shot transfer-learning model that infers and extracts information from free-form multi-modal sources - text, sound, video, shot-splits, speaker-diarization, and face-tracking to create a knowledge-graph using language modelling questionnaire through slot-filling"* [1]. They employ a multimodal transformer that is pre-trained on large video captioning [17] and video question answering [15] datasets. Relations between entities are then obtained using corresponding text prompts.

## 3 PIPELINE DESIGN

The overall pipeline responsible for transforming the raw video input into a semantic graph representation is comprised of several streams, each of which is responsible for a specific aspect. The streams can be grouped into two categories, depending on whether they primarily rely on information coming from the *visual* or the *audio* component of the video.

### 3.1 Visual Streams

All the visual streams use the frames of the video as their primary input.

*3.1.1 Shot Segmentation Stream.* The first stream is concerned with the temporal segmentation of the input video into a series of shots. In our pipeline, we use SceneSeg [19] in order to detect the boundaries of the shots within a video. The stream generates — for every shot — a start and an end timestamp as well as a representative key-frame, to be used by other streams.

*3.1.2 Face Identification Stream.* We rely on face identification to label each shot with the characters it shows. To this end, the stream of key-frames generated by the shot segmentation stream is analyzed using a freely available face detection and identification framework [21, 22]. For every detected face, a feature vector is generated that can be compared to anchor points generated from the example images provided for every character with the dataset using a simple kNN classifier. If the classification probability for a new sample is sufficiently high, it is also added to the list of anchors for the relevant character, thereby providing a more diverse reference during the processing of a movie. The output of this stream is a list of identified characters for every shot, together with the bounding box coordinates for every detection.

*3.1.3 Face Emotion Stream.* Using the same framework as in the face identification stream, we can also estimate the expressed emotion in every face. The face emotion stream uses the detected faces of the previous stream as an input and produces an emotion probability histogram for every face and shot. The emotion categories consist of anger, disgust, fear, happiness, neutrality, sadness, and surprise.

*3.1.4 Action Recognition Stream.* To get an indication of what is happening within a shot, the action recognition stream uses a public implementation [7] of a SlowFast [8] action recognition model trained on the kinetics400 dataset [14]. The stream uses the segmented shots as input and produces the three most likely detected actions for each of them.

*3.1.5 Location Identification Stream.* In order to gain information about the location at which a shot is set, the location identification stream uses two parallel approaches in order to categorize the type of location in general and to relate a shot to one of the known, semantically relevant locations specifically. For the location categorization aspect, a ResNet18 [12] trained on the Places365 [26] dataset is used and the three most likely location classes for every shot are returned as an output. To detect the specific known locations, the stream uses the DELF [18] local interest point descriptor in combination with random sample consensus [9].

### 3.2 Audio Streams

The audio streams primarily rely on the audio information of the video and do not require any direct visual input.

*3.2.1 Speaker Diarization Stream.* The speaker diarization stream aims to identify the audible speaker or speakers in every shot. It uses the audio of each shot as an input and further segments it into speech sequences of the same speaker. For each of these segments, a feature vector is generated that can be used to re-identify a previously heard speaker [4]. Together with the information generated by the face identification stream, these features can be associated to known characters over time. The output of this stream is a series time intervals with an associated label during which a specific speaker is speaking.

*3.2.2 Speech Emotion Recognition Stream.* For every sequence of speech by one specific speaker, the speech emotion recognition stream aims at categorizing the emotion of the voice into happy, sad, angry, and neutral. To do this, it uses classifiers provided by the SpeechBrain [20] framework. The output of this stream is an emotion probability histogram for every time interval.

*3.2.3 Speech Transcription Stream.* Using the same SpeechBrain framework, the speech transcription stream extracts the spoken text for each time interval. The stream outputs the raw detected text for each such interval.

*3.2.4 Speech Sentiment Stream.* Based on the previously transcribed text, the speech sentiment stream estimates for every speech interval if the content of what was said is rather positive or negative. It uses a framework called TextBlob [16] to achieve this. The output is a single number for every interval, estimating the positivity of the content.

*3.2.5 Character Mention Stream.* The character mention stream uses the transcribed text in order to identify in what intervals a character mentions another. The output is then simply a list of mentioned characters for every interval.

### 3.3 Relationship Classification

We combined the visual and audio information by temporally aligning the identified face labels with the those of the speaker diarization stream. Then, we chose the most frequent character for each speaker diarization output to connect the visual, and the audio information. In order to construct the graph from the information generated by the various streams, we first identify the pairs of entities that have a relation out of all possible pairs of the given entities. This is done with a random forest binary classifier that uses all

stream information relevant for the two specific entities as an input. If the classifier determines that the two entities are related, one out of three secondary multi-class classifiers is used to predict the specific relation. The three classifiers are specialized for person to person, person to location, or person to concept relations, respectively. For this relation classification process, a gradient boosted decision tree [10] is used, based on the same information as the previous binary random forest. All classifiers are trained exclusively on the most recent version of the HLVU dataset [5] provided for the challenge.

## 3.4 Query Processing

Based on the output of the various classifiers, we construct a graph that not only contains the detected relations between the entities but also their estimated probabilities. These probabilities can then be used during query processing in order to rank all candidate relations or paths that are assumed to be relevant for a given question. We use the NetworkX [11] framework for graph construction and querying.

# 4 INSIGHTS

In the following, we discuss the insights gained from analysis of the movie-level classifiers, the movie-level knowledge graphs, and the scene-level classifier.

The *person-to-person classifier* predicted only three different relations out of 30 possible different person-to-person relations present in the training set. Compared to the relationship distribution in all the training movies, the amount of different relationships should presumably have been higher. The high frequency of the 'friend of' relation can be explained by the fact, that this is the most common relation in the training dataset. It appears 49 times while the second most often appearing relation in the dataset are 'acquaintance of' and 'parent of'. An overview is shown in Table 1.

**Table 1: Relation distribution for person-to-person relations in test movies**

| relations | frequency |
|---|---|
| friend of | 110 |
| parent of | 4 |
| acquaintance of | 1 |

From 24 different *person-to-location relations* on which we trained, only two different relations appear in the set of predicted person-to-location relations as shown in Table 2. The relation 'socializes at' was also the most common person-to-location relation in the training dataset. There are 140 examples for 'socializes at', while the second relation, which was predicted from the classifier only has 19 different entries in the training dataset.

**Table 2: Relation distribution for person-to-location relations in test movies**

| relations | frequency |
|---|---|
| socializes at | 182 |
| residence of | 32 |

We ran a basic analysis on the predicted knowledge graphs from the test movies. Measurements such as number of different relations, nodes, and edges inspected as well as the graph densities were calculated, which can be found in Table 3.

**Table 3: Knowledge graph analyses test movies**

| movie | rels. | nodes | entities | edges | density |
|---|---|---|---|---|---|
| Calloused Hands | 8 | 9 | 17 | 17 | 23.61% |
| Chained For Life | 3 | 6 | 9 | 6 | 20% |
| Liberty Kid | 4 | 9 | 15 | 15 | 20.83% |
| Like Me | 4 | 8 | 9 | 12 | 21.43% |
| Little Rock | 2 | 16 | 18 | 21 | 8.75% |
| Losing ground | 3 | 12 | 15 | 13 | 9.84% |

A comparison between the number of nodes and the number of available entities showed that none of the knowledge graphs managed to be complete as some entities are missing in all of knowledge graphs. Some of the missing entities are animals as they were not actively tracked in the pipeline. Another issue, which we found by analyzing the knowledge graphs, are the missing locations. For the movie *Calloused Hands* the pipeline misses 6 locations in the graph, while for *Chained For Life* it misses 3 locations, and for *Liberty Kid* it did not even manage to include one of the locations in the knowledge graph. The mean graph density value from all training is 15.0% and the median value is 12.5%. The highest density has a value of 36.67%, while the lowest density has a value of 5.3%. The densities from the predicted knowledge graphs are all in the range of the training movies, which could be a positive sign for the predicted graphs.

**Table 4: Analysis of top 10 scene-level interaction classifier**

| interactions | frequency |
|---|---|
| socializes with | 99 |
| kisses | 94 |
| plays with | 75 |
| arrests | 56 |
| talks to | 44 |
| threatens | 38 |
| compliments | 34 |
| works with | 32 |
| rejects | 29 |
| bullies | 25 |

The *interaction* training dataset contained 53 different interaction and from the test dataset we were able to predict 40 different interactions. We analyzed the ten most often predicted interactions from the scene-level tasks and compared these labels to the training dataset. Table 4 shows the details of the top 10 predicted interactions. In comparison to the movie-level classifiers the most often predicted interaction 'socializes with' was not the interaction, which had the most entries in the interaction training dataset. It is ranked only as the sixth often interaction from the training dataset. The interaction appearing the most often in the training dataset

is the interaction 'talks to'. Therefore, we can assume, that the interaction classifier is not affected by the irregularly distributed interaction labels for the scenes.

## 5 LIMITATIONS AND OPEN CHALLENGES

In the following, we discuss several observed limitations of our approach as well as some inherent in the challenge as such.

*Availability of Training Data.* While the multimedia and machine learning communities have collected large amounts of multimodal data for various purposes over recent years, the specific requirements of this challenge are such that little suitable training data is available that fulfills all requirements. Outside of the comparatively small HLVU dataset [5], parts of which serve as the test set for the challenge and hence come without annotation, only few datasets cover the semantic properties of long-form video. The most related alternatives are the MovieNet [13] and the MovieGraphs [23] datasets, both providing semantic graph annotations for cinematic content. The structure of these graphs is however sufficiently different from both the ones in the HLVU dataset as well as from each other as to make them of limited use for this application.

*Distribution Shift for pre-trained Models.* Similar to other challenge participants, we use models for feature extraction that have been trained on data with a different origin and, hence, generally different properties. Since the challenge dataset is composed of professionally produced video content made for entertainment purposes, it has a different look from visual data that has been produced in a different context, such as for example ImageNet [6], which serves as a training set for many neural networks used as feature encoders. Also, given that no dataset of comparable size with visual properties generally analogous to the challenge dataset is currently available, it is not feasible to quantify the extent of the influence of these differences.

*Face Identification insufficient.* In our current approach, we exclusively rely on faces in order to identify characters in the movies. This choice was mostly made for sake of simplicity, since all relevant characters in the challenge dataset are humans and the example images given do prominently show their faces. There are however obvious limitations to that approach, since it requires these conditions to be true, which is certainly not the case for all video.

*Noise accumulates.* Since in our multi-stream approach, information generated in one stream can be used as input in another, any noise introduced by any of the streams has the potential to be amplified when moving through the entire pipeline. Similarly, information that has been missed by a stream will not be available for any subsequent streams, impeding their performance. We have seen this in Section 4 where some entities have not been detected. This is a fundamental limitation of this iterative approach and cannot easily be circumvented.

## 6 CONCLUSION

In this paper, we presented a multi-stream approach for extracting higher level semantic information from long-form video. Streams can build upon the information extracted in previous streams in order to incrementally obtain more information. The output of the streams can then be aggregated and using several classifiers be transformed into a knowledge-graph representation of the input video.

## REFERENCES


[1] Vishal Anand, Raksha Ramesh, Boshen Jin, Ziyin Wang, Xiaoxiao Lei, and Ching-Yung Lin. 2021. MultiModal Language Modelling on Knowledge Graphs for Deep Video Understanding. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 4868–4872. https://doi.org/10.1145/3474085.3479220

[2] Vishal Anand, Raksha Ramesh, Ziyin Wang, Yijing Feng, Jiana Feng, Wenfeng Lyu, Tianle Zhu, Serena Yuan, and Ching-Yung Lin. 2020. Story Semantic Relationships from Multimodal Cognitions. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4650–4654. https://doi.org/10.1145/3394171.3416305

[3] Matthias Baumgartner, Luca Rossetto, and Abraham Bernstein. 2020. Towards Using Semantic-Web Technologies for Multi-Modal Knowledge Graph Construction. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4645–4649. https://doi.org/10.1145/3394171.3416292

[4] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote.Audio: Neural Building Blocks for Speaker Diarization. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974

[5] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. HLVU: A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*, Cathal Gurrin, Björn Þór Jónsson, Noriko Kando, Klaus Schöffmann, Yi-Ping Phoebe Chen, and Noel E. O'Connor (Eds.). ACM, 355–361. https://doi.org/10.1145/3372278.3390742

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[7] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. 2020. PySlowFast. https://github.com/facebookresearch/slowfast.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 6201–6210. https://doi.org/10.1109/ICCV.2019.00630

[9] Martin A. Fischler and Robert C. Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* 24, 6 (1981), 381–395. https://doi.org/10.1145/358669.358692

[10] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[11] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, Gaël Varoquaux, Travis Vaught, and Jarrod Millman (Eds.). Pasadena, CA USA, 11 – 15.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[13] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. MovieNet: A Holistic Dataset for Movie Understanding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 12349)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 709–727. https://doi.org/10.1007/978-3-030-58548-8_41

[14] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* abs/1705.06950 (2017). arXiv:1705.06950 http://arxiv.org/abs/1705.06950

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi

Tsujii (Eds.). Association for Computational Linguistics, 1369–1379. https://doi.org/10.18653/v1/d18-1167

[16] Steven Loria et al. 2018. textblob Documentation. *Release 0.15* 2 (2018), 269.

[17] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2630–2640. https://doi.org/10.1109/ICCV.2019.00272

[18] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-Scale Image Retrieval with Attentive Deep Local Features. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 3476–3485. https://doi.org/10.1109/ICCV.2017.374

[19] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A Local-to-Global Approach to Multi-Modal Movie Scene Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 10143–10152. https://doi.org/10.1109/CVPR42600.2020.01016

[20] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit. *CoRR* abs/2106.04624 (2021). arXiv:2106.04624 https://arxiv.org/abs/2106.04624

[21] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 23–27. https://doi.org/10.1109/ASYU50717.2020.9259802

[22] Sefik Ilkin Serengil and Alper Ozpinar. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 1–4. https://doi.org/10.1109/ICEET53442.2021.9659697

[23] Paul Vicol, Makarand Tapaswi, Lluís Castrejón, and Sanja Fidler. 2018. MovieGraphs: Towards Understanding Human-Centric Situations From Videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 8581–8590. https://doi.org/10.1109/CVPR.2018.00895

[24] Fan Yu, Dandan Wang, Beibei Zhang, and Tongwei Ren. 2020. Deep Relationship Analysis in Video with Multimodal Feature Fusion. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4640–4644. https://doi.org/10.1145/3394171.3416303

[25] Beibei Zhang, Fan Yu, Yanxin Gao, Tongwei Ren, and Gangshan Wu. 2021. Joint Learning for Relationship and Interaction Analysis in Video with Multimodal Feature Fusion. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 4848–4852. https://doi.org/10.1145/3474085.3479214

[26] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009