

An empirical study of the effect of video encoders on Temporal Video Grounding

Ignacio M. De la Jara[†] Cristian Rodriguez-Opazo[‡] Edison Marrese-Taylor[♣] Felipe Bravo-Marquez[†]
 Department of Computer Science, University of Chile, CENIA & IMFD [†]
 Australian Institute for Machine Learning, University of Adelaide[‡]
 National Institute of Advanced Industrial Science and Technology [♣]
 ignacio.meza@ug.uchile.cl, cristian.rodriguezopazo@adelaide.edu.au
 emarrese@weblab.t.u-tokyo.ac.jp, fbravo@dcc.uchile.cl

Abstract

Temporal video grounding is a fundamental task in computer vision, aiming to localize a natural language query in a long, untrimmed video. It has a key role in the scientific community, in part due to the large amount of video generated every day. Although we find extensive work in this task, we note that research remains focused on a small selection of video representations, which may lead to architectural overfitting in the long run. To address this issue, we propose an empirical study to investigate the impact of different video features on a classical architecture. We extract features for three well-known benchmarks, Charades-STA, ActivityNet-Captions and YouCookII, using video encoders based on CNNs, temporal reasoning and transformers. Our results show significant differences in the performance of our model by simply changing the video encoder, while also revealing clear patterns and errors derived from the use of certain features, ultimately indicating potential feature complementarity.

1. Introduction

Understanding and reasoning about long, untrimmed videos is at the core of Computer Vision (CV). As humans have the ability to intuitively identify relevant moments within videos, the task of Temporal Video Grounding (TVG) appears as a fundamental effort in CV, aiming to develop models that recognise and determine temporal boundaries of action instances in videos [10, 12, 28] using natural language queries [7, 13].

As such, work on the TVG task is extensive and includes a wide variety of approaches and techniques. While the original models were mostly suggestion-based, more recent techniques have aimed at predicting the start and end temporal positions directly, or by regressing them from the input video. The recent advent of transformer-based models [32] has also brought new developments, where the integration of pre-training stages or the direct addition of pre-

trained vision and speech models has led to significant performance improvements.

Despite the large amount of prior work in the TVG task, we find that the role of the video representation has not been consistently investigated so far. Specifically, we observe that prior work has relied on features derived from action classification models such as C3D [30] or I3D [2], with alternatives remaining relatively unexplored. We speculate that this selection bias may lead to model designs that overfit or exploit spurious patterns in these features, without generalising in the long run. We suggest that the increased complexity of recent approaches in search of improved performance can be seen as indirect evidence of this point.

In light of this issue, in this paper we propose a comprehensive empirical study to shed light on the role of video representation in the TVG task. We consider three relevant benchmark datasets, Charades-STA, ActivityNet Captions, and YoucookII, and design a comprehensive framework that allows us to isolate the effect of different video representations by introducing minimal changes in the model architecture. We use a wide variety of pre-trained models to extract video representations, including more than 10 different types of models, resulting in more than 30 sets of extracted features for our data. These include, but are not limited to, well-known CNN and Transformer-based action classifiers. We release features to encourage research in this area. ¹

Our results show that changes in the video representation can lead to substantial performance improvements by keeping the models as is, allowing a “classical” approach [25] to perform better by a large margin by simply changing the video encoder, findings that are consistent with similar observations recently made in the context of query representations [16]. Our experiments reveal clear pitfalls in the use of certain features and uncover complementarity between features, which could lead to further performance improvements if exploited.

¹https://github.com/Mezosky/VideoFeatures_TVG

2. Related Work

Action Classification The task of identifying the action being performed in a video is the cornerstone of current video understanding pipelines. It is essential for the task of temporal video grounding, as it is the current way of encoding videos. Current methods are usually trained via supervised learning, using datasets such as Something-Something [11], Kinetics [2], among others, which contain short video clips of a single action.

Different methods have been proposed to solve the action classification task, the most recent approach based on neural networks could be divided into CNN-based [2, 5, 6, 31], temporal reasoning [19, 33] and transformers [4, 18]. In the case of CNN-based, the methods exploit the ability of CNNs to capture spatial information in images and add various techniques to extract the temporal information from the video. [31] proposed a convolutional block, which is the combination of 2D convolutions followed by a 1D convolution. Such blocks approximate the behaviour of 3D convolutions and allow to capture the temporality from the videos. Knowing that the information in a video is very redundant and not symmetric, [6] proposed to use a slow and a fast track, aiming to capture the spatial and temporal information respectively. Temporal reasoning classifiers aim to capture the temporal information between features extracted from frames using techniques that exploit latent information in videos, such as redundancy (sampling), summarisation (aggregation), and ordering (ranking). Limin et al. [33] use a sampling technique to select features that, after passing through an aggregation mechanism, effectively represent the video. Aiming to create a better representation of the videos, Lin et al. [19] proposed a shift module that applies a shift in part of the temporal channel of the features to exchange information between frames. The advent of transformer-based models has also led to new developments in this task [1, 4]. However, their success in this task is due to their flexibility in capturing spatial information and their inherent ability to model sequential information. [4] presents a multiscale feature hierarchy that combines low-level spatial information processed at early layers with more complex high-level features that capture temporal information processed at deeper layers.

Temporal Video Grounding The task that concerns this work, where we find mainly two types of approaches. On the one hand, we find proposal-based models that, given a query, return a set of candidates that can later be ranked [3, 8, 20, 36, 38]. Recent approaches include both transformer-based models and contrastive losses [24, 27, 35, 37, 39]. On the other hand, we find proposal-free models that try to predict the start and end locations directly from the video span [9, 25].

3. Experimental Framework

Model At the core of our experimental framework lies the TVG model to be used as a pivot for testing video features. Although there are a variety of approaches to our task, we find that not all of them are suitable for our purposes, as replacing the video representations can in many cases lead to substantial changes in the model architecture and/or training. After careful consideration, we selected the proposal-free approach proposed by Rodriguez et al. [25] (TMLGA), which can be summarized in three parts: 1) a text encoder based on GloVe [23] embeddings on top of an LSTM [14], 2) a video encoder, which we discuss in detail below, and, 3) a localization module that combines information from both modalities using a dynamic filter [15] and predicts the start and end points of the segment. We consider this model to be a “*classic*” approach, as it has been widely adopted as a baseline by recent work, allowing for meaningful comparisons with a wide variety of approaches. We also note that TMLGA relies on a rather simple approach to query representation, which we believe allows for a cleaner result in terms of the contribution of the language encoder. Finally, we point out that TMLGA has an official implementation that has been publicly released, which we believe is essential to facilitate the replication of the original results and ultimately ensure the reproducibility of our experiments.

Video Representation Following previous work, our pivot TMLGA model uses pre-trained action classification models to obtain a sequence of vectors to represent a given video. Critically, these input vectors are just passed through one projection layer before being fed into the localization module, which ensures that the amount of changes introduced to the model is minimized for our experiments. Although the original implementation relies on I3D [2] we instead consider a wide selection of pre-trained models for feature extraction, which we propose to group into three main categories. First, we test CNN-based video classification approaches, including C2D [30], I3D, SlowFast [6], X3D [5] and Non-Local variations in one classifier [34]. For these models, we modify the classification head of the last ResNet or ConvNet used, extracting the representation generated before dropout, final projections and classification layer of the network. Secondly, we look at Transformer-based action classification models, including MViT [4], MViT2 [18], and Rev-MViT [21]. Similarly to the CNN-based case, for these models we represent the video using the vectors of the last layer of the transformer before the linear classifier. Finally, we study temporal reasoning models such as TSM [19], which provide a different angle in tackling the action classification task. In this model, the video data is divided into segments, each containing 8 buckets, where each bucket consists of 8 consecutive frames. This segmentation approach creates smaller

video subsequences, allowing the temporal network to effectively permute frames and capture their temporal dependencies. Similarly to the other cases, the last projections of the temporal network are discarded to obtain the video feature representation.

Following Rodriguez et al. [25], the video encoding procedure is done offline, performing rescaling to 320x240 when necessary, and sampling at 25 fps. Our implementation is based on *decord*², *pyslowfast*³ and the official code release for TSM [19]⁴. Each video is sequentially traversed in buckets of a size given by the frame rate of the backbone model. In cases where the total number of extracted frames is not divisible by this number, we fill the last bucket by repeating the last frame.

Datasets We consider three well-known challenging benchmarks for the task. *Charades-STA* [7, 29], *ActivityNet-Captions* [17], and *YouCookII* [40, 41]. We use the pre-defined train and test sets. Each dataset with its own challenges, from more ambiguous spatio-temporal information to very long videos.

Evaluation To assess the effectiveness of our extracted features, we conduct two kinds of analysis. First, a quantitative one based on metrics proposed in [7], namely the recall with thresholds (α) of 0.5 and 0.7 for the temporal intersection over union (tIoU). Secondly, we provide an in-depth qualitative analysis of the model predictions, focusing on the biases presented when training with different features.

4. Results

Quantitative Analysis Table 1 summarizes the results of our experiments, showing the performance of the TMLGA model trained with different features. As can be seen, in the case of the Charades-STA dataset, it is clear that the Transformer-based features give promising results, while our I3D features still perform poorly. This is in contrast to the original results, where the I3D features were only able to deliver 2.45 points over our results. We attribute this difference to the pre-training dataset, as the original features were pre-trained on Charades, which may have led to overfitting. The proposed features, when pre-trained on the K400 dataset, show competitive performance compared to the original results. This allows for unbiased selection and avoids the need for time-consuming training on the task-specific dataset, thereby accelerating the development of architectures for localization tasks.

On ActivityNet, we note two interesting observations. First, we find that some CNN-based features can outperform more recent Transformer-based models, with differ-

²<https://github.com/dmlc/decord>

³<https://github.com/facebookresearch/SlowFast>

⁴<https://github.com/mit-han-lab/temporal-shift-module#pretrained-models>

Encoder	Model		Charades		ActivityNet		YouCookII	
	f	Pretrain	0.5	0.7	0.5	0.7	0.5	0.7
MViT	8	K400	50.27	31.32	30.64	17.50	27.78	15.21
MViT-v2	16	K400	48.17	30.27	23.80	11.87	25.40	13.95
Rev-MViT	16	K400	49.89	32.31	28.94	15.97	23.71	13.00
X3D M	16	K400	43.60	27.96	26.39	14.14	22.97	12.80
X3D S	13	K400	41.37	25.16	30.23	17.54	23.00	12.60
SlowFast	8	K400	38.82	24.22	<u>23.34</u>	<u>11.70</u>	21.39	11.57
SlowFast	16	SS V.2	39.76	23.90	30.15	17.51	21.16	11.05
SlowOnly	8	K400	36.16	20.19	23.92	12.11	11.31	6.24
I3D NLN	8	K400	38.63	24.33	24.04	12.28	20.76	11.60
C2D	8	K400	<u>02.63</u>	<u>00.97</u>	24.33	12.39	<u>08.85</u>	<u>03.92</u>
TSM	8	K400	45.35	23.20	31.81	18.06	24.74	13.29
Mean	-	-	39.51	23.98	27.37	14.90	21.01	11.38
STD	-	-	13.41	06.65	07.04	05.46	06.76	03.37
I3D [26]	8	Charades	52.02	33.74	-	-	-	-
I3D [26]	8	K400	-	-	33.04	19.26	20.65	10.94

Table 1: Summary of our experimental results, where f denotes the frame rate used to extract features, while K400 and SS V.2 stand for the Kinetics-400 and Something-Something V.2 datasets, respectively. In the table, the best results for each dataset are indicated in **bold**, while the worst results are underlined.

ences of up to 7 points in the 0.7 band. We also notice that the overall performance of the features shows less variation, suggesting that the query representation may be particularly relevant. This is consistent with the data showing that ActivityNet has relatively more complex queries in terms of vocabulary size (748 vs. 9,744 tokens) and length (7.2 vs. 13.48 tokens per query).

The results on YouCookII show that this dataset remains the most visually challenging, which we suspect is partly due to the length of the videos, with an average of around 5 minutes. On this dataset, Transformer-based models again achieve the best performance, with results similar to complex architectures such as DORi [26], which make explicit use of spatial information.

Overall, our results highlight the importance of careful feature selection in the TVG task, as we see that a simple change can lead to significant performance gains of up to 5 points. Ultimately, we see how selecting features without a thorough understanding of their impact on the dataset can lead to stagnation in development and suboptimal performance. Further research is suggested to investigate the effect of different features and to explore their applicability to other video and speech tasks.

Qualitative Analysis To visualize the output of the model when given different features, we plot the 0-1 normalized predicted time intervals for each query, similar to [22]. We generate one plot for correct intervals and another for incorrect predictions, based on a given α of 0.7 tIoU band.

As shown in Figure 1, we see that in the case of Charades-STA, different features produce two points of correct predictions, represented by the lower-left and upper-right accumulation points. We see that in many cases, the

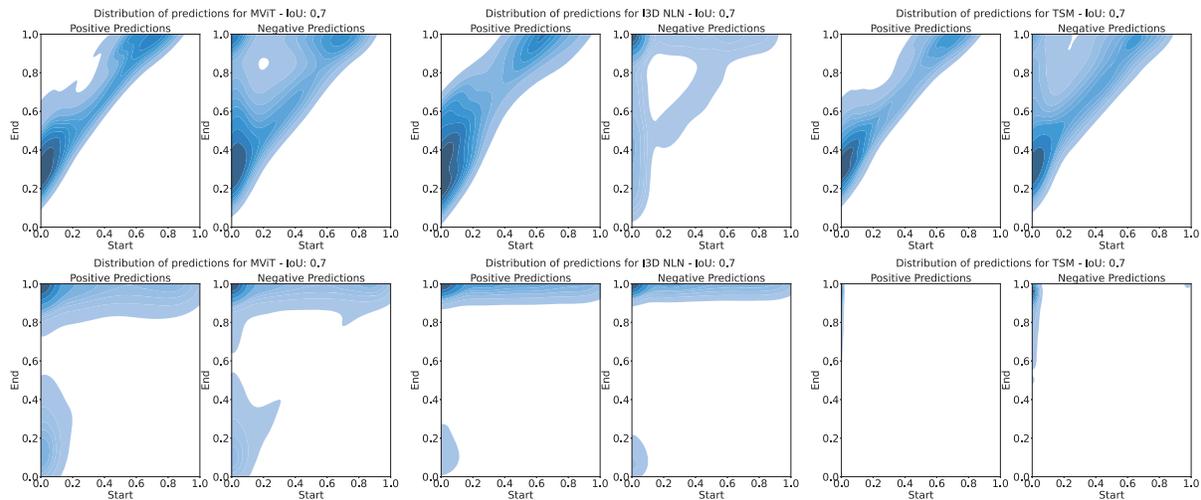


Figure 1: The distribution of normalized temporal predictions with extracted features. The x-axis represents the prediction start time, while the y-axis represents the prediction end time. The graph shows normalized predictions for the Charades dataset (top) and the ActivityNet dataset (bottom), using features from MViT (left), I3D NLN (center) and TSM (right).

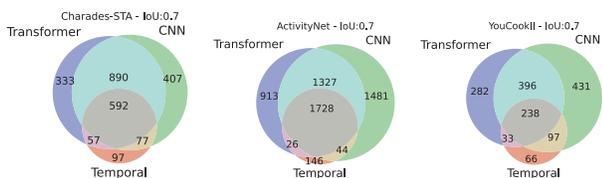


Figure 2: Overlap of the correct predictions, in terms of query-video pairs for $\alpha = 0.7$, in Charades-STA (left), ActivityNet (center) and YoucookII (right) benchmarks. We group features based on their nature.

model tends to produce predictions that cover the entire video length, despite the lack of such examples in the training data. We also see the extent to which the predictions vary between features, with clearly different accumulation zones for each case. This indicates that there are significant differences in positive and negative predictions between models on this dataset, suggesting orthogonality between features. For ActivityNet, we observe that the model tends to fall into degenerate solutions in all cases, making predictions for the entire video length. We believe this may be due to the presence of such annotations in the training data. Despite this, similarly to Charades, different accumulation points are evident in the predictions, indicating potential orthogonality between the features.

Finally, as shown in Figure 2, we plot Venn diagrams to visualize the overlap of correct predictions. We see that there are exclusive predictions based on feature type across the datasets, supporting the visualization of the differences seen in the bias plots. We believe that these results further indicate the presence of orthogonality between the different features, suggesting that the use of different features could be complementary to solving the localization task.

5. Conclusions

In this paper, we have conducted a comprehensive analysis of video features in the context of the video grounding task. Our research shows that using features with pre-trained encoders from datasets different from the target application can produce similar results to using a fine-tuned encoder specific to the application dataset. This approach effectively mitigates potential network overfitting and avoids biases arising from assumptions about task proficiency. Our investigation has also revealed exclusive predictive patterns among video features, suggesting the potential for improved predictions through their integration into a unified network. Furthermore, we have shown that varying the features can have a significant impact on localisation, leading to performance improvements comparable to more complex networks that explicitly consider spatial components in their inputs. These results provide valuable insights into how the spectrum of feature diversity and encoder pre-training can be exploited to advance the field of video analysis and improve the field of temporal localisation tasks. Looking ahead, a notable avenue for future work lies in the application of these features within modern architectures. In addition, it is anticipated that the inherent orthogonality of these features will be exploited to build more robust frameworks tailored to the task of temporal localization.

Acknowledgements

Ignacio Meza and Felipe Bravo-Marquez were supported by ANID Millennium Science Initiative Program Code ICN17_002 and the National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 213–229, Cham, 2020. Springer International Publishing. 2
- [2] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 2
- [3] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. *AAAI*, 2019. 2
- [4] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 2
- [5] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition, 2020. 2
- [6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 2
- [7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 1, 3
- [8] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *WACV*, 2019. 2
- [9] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2
- [10] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video Action Transformer Network. In *CVPR*, pages 244–253, 2019. 1
- [11] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haebel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2
- [12] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1
- [13] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1
- [14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2
- [15] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 2
- [16] Erica Kido Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi, Hideki Nakayama, and Yusuke Miyao. Towards parameter-efficient integration of pre-trained language models in temporal video grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13101–13123, Toronto, Canada, July 2023. Association for Computational Linguistics. 1
- [17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-Captioning Events in Videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017. 3
- [18] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MVITv2: Improved Multiscale Vision Transformers for Classification and Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 2
- [19] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 2, 3
- [20] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 15–24. ACM, 2018. 2
- [21] Karttikeya Mangalam, Haoqi Fan, Yanghao Li, Chao-Yuan Wu, Bo Xiong, Christoph Feichtenhofer, and Jitendra Malik. Reversible vision transformers, 2023. 2

- [22] Esa Rahtu Mayu Otani, Yuta Nakahima and Janne Heikkilä. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In *The British Machine Vision Conference (BMVC)*, 2020. 3
- [23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2
- [24] Cristian Rodriguez, Edison Marrese-Taylor, Basura Fernando, Hiroya Takamura, and Qi Wu. Memory-efficient temporal moment localization in long videos. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1901–1916, 2023. 2
- [25] Cristian Rodriguez, Edison Marrese-Taylor, Fateh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2464–2473, 2020. 1, 2, 3
- [26] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1079–1088, 2021. 3
- [27] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hiroya Takamura, and Qi Wu. Locformer: Enabling transformers to perform temporal moment localization on long untrimmed videos with a feature sampling approach. *arXiv preprint arXiv:2112.10066*, 2021. 2
- [28] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1
- [29] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 510–526, Cham, 2016. Springer International Publishing. 3
- [30] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features With 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 1, 2
- [31] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1
- [33] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2740–2755, 2019. 2
- [34] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks, 2018. 2
- [35] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2613–2623, June 2022. 2
- [36] Huijuan Xu, Kun He, L Sigal, S Sclaroff, and K Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019. 2
- [37] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-Stage Aggregated Transformer Network for Temporal Language Localization in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12669–12678, 2021. 2
- [38] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-Scale 2D Temporal Adjacency Networks for Moment Localization with Natural Language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2
- [39] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly Supervised Temporal Sentence Grounding With Gaussian-Based Contrastive Proposal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15555–15564, 2022. 2
- [40] Luowei Zhou, Nathan Louis, and Jason J. Corso. Weakly-Supervised Video Object Grounding from Text by Loss Weighting and Object Interaction. *arXiv:1805.02834 [cs]*, May 2018. 3
- [41] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards Automatic Learning of Procedures From Web Instructional Videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, Apr. 2018. 3