W-ART: ACTION RELATION TRANSFORMER FOR WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

Mengzhu Li[1] *Hongjun Wu*[1] *Yongcheng Liu*[2] *Hongzhe Liu*[1⋆] *Cheng Xu*[1] *Xuewei Li*[1]

[1] Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China

ABSTRACT

Weakly-supervised temporal action localization (WTAL) is a long-standing and challenging research problem in video signal analysis. It is to localize the action segments in the video given only video-level labels. The key to this task is understanding how the diverse actions interact. In this paper, we propose W-ART, a relation Transformer to explicitly capture the relationships between action segments. We devise a new effective Transformer architecture and construct new training loss functions for WTAL. Further, we propose a dedicated query mechanism to satisfy the different feature preferences between classification and localization. Thanks to these designs, our W-ART can accurately localize the diverse actions even in weakly-supervised setting. Extensive evaluation and empirical analysis show that our method outperforms the state of the arts on two challenging benchmarks, Charades and THUMOS14.

Index Terms— Weakly-supervised Temporal Action Localization, Long-range Temporal Segment Dependency, Relationship Transformer, Weakly-supervised Query Mechanism

1. INTRODUCTION

Temporal action localization (TAL) is a challenging task in video understanding and is widely applied for quickly localize action segments with varying temporal extents. Jin *et al.*[1] and SRF-Net[2] show that TAL has made a huge contribution to video signal processing. This task aims to localize the start and end timestamps in a video for each semantic action, and it can be achieved under supervised or weakly-supervised setting. For supervised setting, training videos are manually annotated with the frame-level timestamps and the label of each action, resulting in the waste of time and the non-diversity of actions. In contrast, weakly-supervised methods classify and localize actions with only video-level labels (*e.g.* play tennis) that indicate whether actions are in the video, thus this setting provides a labor-saving but more challenging solution. Here we tackle this task under weakly-supervised setting.

This work was supported, the National Natural Science Foundation of China (Grant No. 61871039, 62171042), the Academic Research Projects of Beijing Union University(No. ZB10202003, ZK40202101, ZK120202104).
*corresponding author: liuhongzhe@buu.edu.cn

Fig. 1. The relation of long-range actions.

Without frame-level annotations, weakly-supervised system employs the similarity of the same action to determine its full extent and the dissimilarity of different actions to classify labels. Thus, W-TALC[3] and Autoloc[4] use co-activity similarity loss[5] with feature similarity for location and use multi-instance learning[6] loss with feature dissimilarity for classification. However, they do not model the relationships between long-range temporal segments. For example, in Fig.1, 'open wardrobe' and 'close wardrobe' share information but are separated by long-segment action 'fold clothes'. Therefore, when generating proposals of 'close wardrobe', above methods are unlikely to capture the dependency with the information of 'open wardrobe'.

Due to the advantages of local connectivity and translation equivariance, 3D convolutions and graph convolutions have been used to model segments' relations (Cordonnier *et al.*[7], G-TAD[8] and PGCN[9]). But these convolutions are designed to capture short-range information, rather than long-range dependencies that beyond the convolutional receptive field. Though D3d[10] extends the receptive field, it is insufficient in capturing long-range dependencies by aggregating shorter-range information. Instead of convolution, Xu *et al.*[11] uses a recurrent neural network to capture temporal segments' relations. However, this method can not ensure that all temporal segments are treated similarly.

Since videos and sentences are both sequential, Video Understanding shares similarities with Natural Language Processing(NLP). Transformer[12] proposed for capturing long-range relation with self-attention makes great progress in NLP, and has been largely applied in visual field (ViT[13] and DETR[14]). **The meaning of a word can be understood precisely only by relating it to other words in the sentence, similarly, action segments need to be contextualized with the rest of the video.** Following such motivation, we explore Transformer in weakly-supervised temporal action localization (WTAL), for further advancing this area.

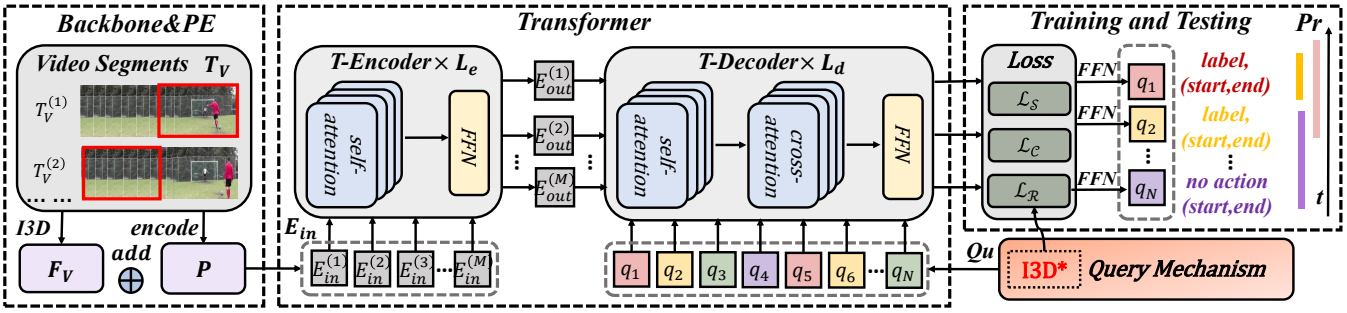

Fig. 2. The overview of the proposed W-ART.

Inspired by object detection, we propose a novel model: W-ART (in Fig.2) to enhance the relationships among long-range temporal segments with common content. Our contributions are highlighted as follows: (1) a kernel query mechanism is designed for WTAL, which solves the issue of matching the temporal locations of video segments; (2) a global matcher loss is proposed for weakly-supervised training, which achieves unique predictions via bi-partite matching; and (3) a reconstruction strategy is proposed, which preserves the feature discrimination by frozen training backbone. W-ART outperforms the state of the arts on two challenging benchmarks: THUMOS14[15] and Charades[16].

2. METHOD

Overview. In Fig.2, given a video containing actions, we aim to localize each action timestamps and predict its classification. In ***Backbone&PE***, we use I3D to extract the feature F of video segments and encode video sequentially temporal information as position encoding (PE) P. With an element-wise sum of F and P, E_{in} is obtained for ***Transformer*** that consists of T-Encoder and T-Decoder. Both of them contain self- or cross- attention mechanisms for capturing the relations of their inputs. Due to lacking known timestamps information, we propose ***Query Mechanism***. A processed action query set Qu is generated and fed into T-Decoder, making the Transformer localize timestamps for WTAL task. In ***training***, we input videos and video-level labels into the model with a matcher loss, which includes the $\mathcal{L}_S$ of localization, the $\mathcal{L}_C$ of classification and the $\mathcal{L}_R$ of feature reconstruction. In ***testing***, we input videos without labels, and the output is a set of action queries with their timestamps, and labels. Adjacently temporal action queries with the same label are merged as a segment and get the final prediction Pr.

Two aspects are in the following: 1) Model Description; and 2) Training and Testing.

2.1. Model Description

Backbone&PE. In Fig.2, to obtain I3D[17] feature F_V of a video V containing t frames, we divide the t into segments of 8 frames with an overlap of 4 frames, resulting in segments $T_V = \{T_V^{(1)}, T_V^{(2)}, ...\}$. As the size of T_V varies with the duration of videos, we set $M = max\{T_{V_1}, T_{V_2}, ...\}$ to unify the dimension. $F_V \in \mathbb{R}^{M\times(D=2048)}$ consists of the feature of RGB ($\mathbb{R}^{M\times1024}$) and Flow ($\mathbb{R}^{M\times1024}$) with a cascade sum. Similar to the token in NLP, we prepare a learnable encoding set $P \in \mathbb{R}^{M\times D}$ for the sequence of feature F, as to retain temporal positional information of each segment. But P carries no temporal information initially. Thus P is randomly initialized at different temporal positions and added to the F_V with an element-wise for training. Finally, we get the input of T-encoder $E_{in} \in \mathbb{R}^{M\times D}$.

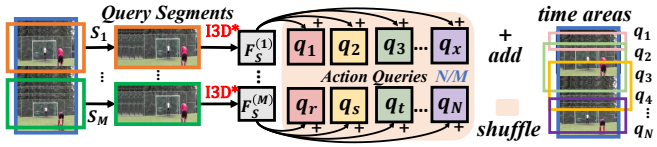

Fig. 3. Query Mechanism.

Query Mechanism. Due to weakly-supervised setting, we randomly crop M segments shown in Fig.3 from a video as query segments $\{S_1, ..., S_M\}$ and record their timestamps as ground truth. The feature of segments $F_S = \{F_S^{(1)}, .., F_S^{(M)}\} \in \mathbb{R}^{M\times D}$, is re-extracted by I3D* (in Sec.2.2-I3D*). We randomly generate N (much larger than M) time areas containing start and end timestamps which are encoded as $Qu = \{q_1, q_2, ..., q_N\} \in \mathbb{R}^{N\times D}$, called action queries. After we randomly assign each $F_S^{(i)}$ to N/M action queries, Qu is input into the T-Decoder. Our purpose is to find the best match action query (timestamps) for each query segment. At last, an action query can precisely localize timestamps of the assigned query segment, through continuously generating more suitable timestamps in training (in Sec2.2). To simulate implicit group assignment among action queries, we randomly shuffle the permutation of all action queries.

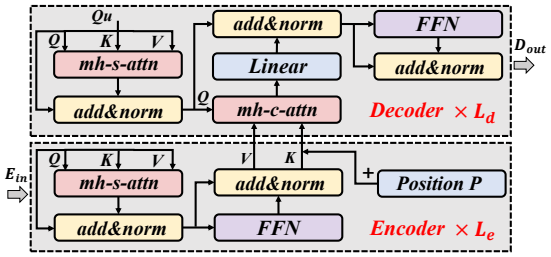

Fig. 4. Transformer Architecture.

Transformer. In Fig.4, Transformer contains T-Encoder and T-Decoder modules. T-Encoder consists of a set of L_e encoder blocks. The l-th encoder block is E_l for $l \in [1, L_e]$. Input and output of E_l are defined as $E_{in}^{(l)}$ and $E_{out}^{(l)}$. Thus,

as a cascade structure, $E_{in}^{(l+1)}$ is equal to $E_{out}^{(l)}$. Each encoder block with standard Transformer architecture performs multi-head self-attention. We add E_{in} to the input of each attention layer. In self-attention, the l-th block's input $E_{in}^{l} \in \mathbb{R}^{M\times D}$, is regarded as query(Q), key(K) and value(V). Single head self-attention is defined as

$$s\_attn_i^{(l)} = softmax(E_{in}^{l} \times (E_{in}^{l})^T/\sqrt{D})E_{in}^{l} \quad (1)$$

where $s\_attn_i^{(l)} \in \mathbb{R}^{M\times \frac{D}{h}}$. After *softmax*, the attention map is obtained as $\mathbb{R}^{M\times M}$. Multi-head self-attention is the concatenation of h single-head attentions and is defined as

$$mh\_s\_attn^{(l)} = [attn_1^{(l)}; ...; attn_h^{(l)}] \quad (2)$$

where mh_s_attn $\in \mathbb{R}^{M\times D}$, since each head is independent and contains $1/h$ of the final result. More details of FFN and add&norm are described in [12].

T-Decoder consists of a set of L_d decoder blocks and performs multi-head self- and cross- attentions. Similar to the multi-head self-attention of T-Encoder, action queries in set Qu are added to the input of each single-attention layer as (key, query, value). We get the output of multi-head self-attention, $\tilde{Qu}$=$[attn_1^{(l)}; ...; attn_h^{(l)}] \in \mathbb{R}^{N\times D}$. For multi-head cross-attention, the output of T-Encoder $E_{out}^{(L_e)}$ is equal to the input $D_{in}^{(1)} \in \mathbb{R}^{M\times D}$ of T-Decoder, which provides value(V) . With an element-wise sum of position P (introduced in Sec2.1) and value V, we get the $\tilde{D}_{in}^{(1)} \in \mathbb{R}^{M\times D}$ as the key(K). Single cross-attention is defined as c_attn_i=$softmax(\tilde{Qu} \times \tilde{D}_{in}^{(1)T}/\sqrt{D}) \times D_{in}^{(1)}$. where c_attn_i is $\mathbb{R}^{N\times D}$. After *softmax*, the attention map is $\mathbb{R}^{N\times M}$. Finally, $mh\_c\_attn = [c\_attn_1^{(l)}; ...; c\_attn_h^{(l)}] \in \mathbb{R}^{N\times D}$.

2.2. Training and Testing

In training, videos with video-level labels are sent to the model. Global matcher loss includes three losses: $\mathcal{L}_{\mathcal{S}}$, $\mathcal{L}_{\mathcal{R}}$, and $\mathcal{L}_{\mathcal{C}}$. Action query set Qu and Qu (after T-Decoder) are sent to $\mathcal{L}_{\mathcal{S}}$. In each epoch, we find the best match action query for query segment. With training, random action queries are likely to be generated in time areas of ground-truth, and the final best match action query is nearly as same as the ground-truth. Therefore, Transformer can localize any segment's timestamps. Query segments' feature F_S and the final prediction Pr are input to $\mathcal{L}_{\mathcal{R}}$, which preserves the feature discrimination for classification (details in I3D*). $\mathcal{L}_{\mathcal{C}}$ is applied for classification, thus it is employed on different videos. Generally, we find segments containing actions with $\mathcal{L}_{\mathcal{C}}$ and $\mathcal{L}_{\mathcal{R}}$, then localize them with Transformer and $\mathcal{L}_{\mathcal{S}}$. Finally, adjacently temporal action queries with the same label are merged as a segment to get the final prediction Pr. In testing, untrimmed videos without labels are sent to the model for predicting their action classifications and localizations.

Formally, $\mathcal{A} = \{a^{(i)}\}_{i=1}^{M}$ is ground truth set where $a^{(i)} = (c^{(i)}, s^{(i)}, r^{(i)})$ and $\tilde{\mathcal{A}} = \{\tilde{a}^{(i)}\}_{i=1}^{N}$ is prediction set. N is larger than M, thus at least $(N\text{-}M)$ action queries are predicted as no action. We compute the same match cost between the prediction $\tilde{a}^{(\tilde{\sigma}(i))}$ and ground-truth $a^{(i)}$ using Hungarian algorithm[18] as matcher loss, where $\tilde{\sigma}(i)$ is the index of $a^{(i)}$ computed by the optimal bipartite matching. Hungarian loss for all matched pairs is defined as

$$\mathcal{L}_{\mathcal{H}} = \sum_{i=1}^{N}\{\mathbb{I}_{\{c^{(i)}\neq\emptyset\}}L_C(\tilde{P}_{\tilde{\sigma}(i)}) + \mathbb{I}_{\{c^{(i)}\neq\emptyset\}}\mathcal{L}_{\mathcal{S}}(s^{(i)}, \tilde{s}^{\tilde{\sigma}(i)})) + \mathbb{I}_{\{c^{(i)}\neq\emptyset\}}\mathcal{L}_{\mathcal{R}}(r^{(i)}, \tilde{r}^{\tilde{\sigma}(i)})\}. \quad (3)$$

where $L_C(\tilde{P}_{\tilde{\sigma}(i)})$=$\tilde{P}_{\tilde{\sigma}(i)}(c^{(i)})$ is the probability of class $c^{(i)}$ for prediction. Video-level labels are used to classify actions. We average the top k per class to get a c dimension video-level prediction. It is defined as:

$$\mathcal{L}_{\mathcal{C}} = -1/n\sum_{j=1}^{n}\sum_{i=1}^{c} c_i^j log(\tilde{P}_{\tilde{\sigma}(i)}^j). \quad (4)$$

$\mathcal{L}_{\mathcal{S}}$ represents segment loss that measures proximity in the timestamps $s^{(i)}$=$[t_s^{(i)}, t_e^{(i)}]$ of actions. The segment loss is defined as a weighted combination of a L_1 loss (sensitive to the durations of instances) and an IoU loss (invariant to the durations of instances) between the prediction and ground-truth. It is expressed as

$$\mathcal{L}_{\mathcal{S}} = \lambda_{iou}\mathcal{L}_{iou}(s^{(i)}, \tilde{s}^{\tilde{\sigma}(i)}) + \lambda_{L_1}\|s^{(i)} - \tilde{s}^{\tilde{\sigma}(i)}\|_1. \quad (5)$$

Feature reconstruction: I3D*. Temporal action localization is the coupling of classification and localization, where these two tasks have different feature preferences[19]. We propose a feature reconstruction loss $\mathcal{L}_{\mathcal{R}}$ to preserve classification feature during localization training. The motivation is to preserve the feature discrimination extracted by I3D backbone after passing feature to Transformer. $\mathcal{L}_{\mathcal{R}}$ is the mean squared error among the 2-normalized segment features extracted by the I3D backbone, which is defined as

$$\mathcal{L}_{\mathcal{R}}(r^{(i)}, \tilde{r}^{\tilde{\sigma}(i)}) = \|\frac{r^{(i)}}{\| r^{(i)} \|_2} - \frac{\tilde{r}^{\tilde{\sigma}(i)}}{\| \tilde{r}^{\tilde{\sigma}(i)} \|_2}\|_2^2. \quad (6)$$

We freeze backbone I3D and propose feature reconstruction I3D* to preserve feature discrimination for classification.

3. EXPERIMENT

Dataset and Implementation. THUMOS14[15] has annotations for 20 classes, with 200/211 untrimmed validation/test videos. We set the max number of segments M=256 and action queries N=300, and do not use dropout. Charades[16] contains 9.8K videos, among which about 8K for training and 1.8K for validation. We use dropout with default probability 0.1. We set the max number of segments M=64 and action queries N=100. Totally, we train the model by AdamW optimizer[20] with a learning rate of 1e-5 and a weight decay of 1e-5 for 3000k steps. We set epoch=15 and reduce the learning rate by a factor of 10 after 2000k steps. Hyperparameters in the loss λ_{L_1} and λ_{iou} are set to 5 and 3. We set base model dimension in the transformer as 512 and set the number of T-Encoder and T-Decoder layers as 4 with 8 attention heads. All learnable weights are initialized using Xavier initialization.

Comparison with state-of-the-art (SOTA). We compare the performance of our W-ART with the SOTA methods. We use mean average precision (mAP) as the metric to evaluate our model. Table.1 shows that the our W-ART achieves to 0.6% improvement over SOTA for UNT feature and 0.7% for I3D

feature in THUMOS14. Table.2 shows the comparisons with SOTA methods on Charades with 0.5% improvement. Overall, results clearly show that our proposed method W-ART outperforms the SOTA methods.

Table 1. Comparison with SOTA (THUMOS14). We report the mean average precision at different intersection over union thresholds (mAP@tIoU) for tIoU$\in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. $\uparrow$ indicates higher is better.

	Methods	mAP@IoU ↑				
		0.3	0.4	0.5	0.6	0.7
UNT	STPN[21](2018)	31.1	3.5	16.2	9.8	5.1
	W-TALC[3](2018)	32.0	26.0	18.8	10.9	6.2
	AutoLoc[4](2018)	35.8	29.0	21.2	13.4	5.8
	CleanNet[22](2019)	37.0	30.9	23.9	13.9	7.1
	W-ART*	37.5	31.4	**24.5**	14.2	7.6
I3D	STPN[21](2018)	35.5	25.8	16.9	9.9	4.3
	W-TALC[3](2018)	40.1	31.1	22.8	14.5	7.6
	MAAN[23](2019)	41.1	30.6	20.3	12.0	6.9
	BaSNet[24](2020)	44.6	36.0	27.1	18.6	10.4
	DGAM[25](2020)	46.8	38.2	28.8	19.8	11.4
	W-ART*	50.7	41.1	**29.5**	21.2	12.7

Table 2. Comparison with state-of-the-art (Charades). We report mAP to compute using setting in [16].

Methods	mAP↑
Two-stream[16](2016)	8.9
Two-stream + LSTM[16](2016)	9.6
R-C3D[26](2017)	12.7
SSN[27](2017)	16.4
I3D baseline[28](2019)	17.2
TGM[29](2018)	22.3
Mavroudi *et.al* [30](2020)	23.7
W-ART*	**24.2**

Table 3. Ablation Study. We use mAP as metrics. E and D mean the number of T-Encoder/-Decoder. H represents the head of Transformer. Q is on behalf of action query.

Dataset	E	D	mAP	H	mAP	Q	mAP
Thumos	4	2	29.0	8	29.3	150	25.2
	4	4	29.3	4	27.3	300	29.2
	2	4	29.1	2	24.9	900	29.3
Charades	4	2	24.1	8	24.2	30	20.1
	4	4	24.2	4	21.9	50	24.4
	2	4	23.9	2	20.1	100	24.4

Ablation Study. Table.3 shows ablation study results in the number of Transformer layers, heads, and action queries. (1) Impact on the number of Layers in T-Encoder/Decoder. While an increasing number of layers results in training time, we do not observe much difference in the performance of the model with increasing depth of the Transformer components. (2) Effect on the number of heads in Transformer. The results suggest a slight improvement with more heads in the transformer. (3) Influence on the number of action queries. Intuitively, the large size of action queries can make a better result though requiring longer training time. However, minor improvement with more nodes in the action queries. Results suggest that when the size of the action queries is reduced, the localization performance of our model degrades. We provide the mAP values to average over the various intersection-over-union thresholds (tIoU) for datasets. In Fig.5, we compare frozen I3D and feature reconstruction to illustrate their importance. Lines a-d show the results on Thumos14 and e-h are in Charades. A-d consist of both modules, a frozen I3D module, no modules, and a feature reconstruction module. E-h are similar to a-d. Obviously, d with only reconstruction module performs worse than c without these two modules, as same as h is worse than g. It confirms that freezing I3D is essential to training. Besides, it also proves the feature reconstruction weaken the effect without frozen I3D.

Visualization. In Fig.6, we visualize the predictions of the model with the sample (in Sec.Introduction). The visualization indicates that our model is able to predict the correct number of actions as well as their categories with minimal errors in start and end timestamps. However, other models can not effectively identify the action of closing wardrobe. Due to that video content around the start and end timestamps in this action does not contain enough information pertaining of the action, such as the common content 'wardrobe'.

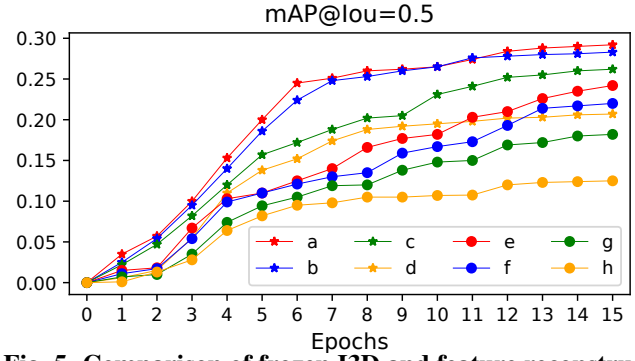

Fig. 5. Comparison of frozen I3D and feature reconstruction. Record the mAP in each epoch with different modules.

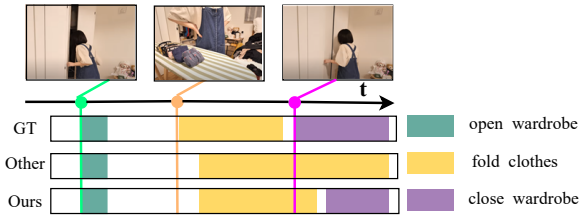

Fig. 6. Qualitative Results. Visualization of ground truth, other methods and ours. GT means ground-truth.

4. CONCLUSION

In this paper, a novel framework named W-ART for weakly-supervised temporal action localization (WTAL) is proposed. It consists of a new effective Transformer architecture and several new training loss functions. Besides, we design a dedicated query mechanism for WTAL to satisfy the different feature preferences between classification and localization. Extensive results demonstrate the effect of our approach. To our knowledge, it is the first time that Transformer-based architecture has been considered for WTAL.

5. REFERENCES

[1] C Jin et al., "Regression before classification for temporal action detection," in *ICASSP*. IEEE, 2020.

[2] R Ning et al., "Srf-net: Selective receptive field network for anchor-free temporal action detection," in *ICASSP*. IEEE, 2021.

[3] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *ECCV*, 2018.

[4] Z Shou et al., "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *ECCV*, 2018.

[5] S Paul, S Roy, and A Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *ECCV*, 2018.

[6] Zhi-Hua Zhou, "Multi-instance learning: A survey," *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 2004.

[7] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi, "On the relationship between self-attention and convolutional layers," *arXiv preprint arXiv:1911.03584*, 2019.

[8] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem, "G-tad: Sub-graph localization for temporal action detection," in *CVPR*, 2020.

[9] Runhao Zeng et al., "Graph convolutional networks for temporal action localization," in *ICCV*, 2019.

[10] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar, "D3d: Distilled 3d networks for video action recognition," in *ICCV*, 2020.

[11] Y Xu, C Zhang, et al., "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," in *AAAI*, 2019.

[12] N Vaswani, Aand Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, Lukasz Kaiser, and I Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[13] A Dosovitskiy, L Beyer, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, 2020.

[14] N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov, and S Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020.

[15] H Idrees, A R Zamir, Y Jiang, A Gorban, I Laptev, and M Sukthankar, Rand Shah, "The thumos challenge on action recognition for videos "in the wild"," *CVPR*, 2017.

[16] GA Sigurdsson et al., "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *ECCV*, 2016.

[17] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition," *A new model and the kinetics dataset. CoRR, abs/1705.07750*, vol. 2, no. 3, pp. 1, 2017.

[18] R Stewart, M Andriluka, and AY Ng, "End-to-end people detection in crowded scenes," in *CVPR*, 2016.

[19] Haichao Zhang and Jianyu Wang, "Towards adversarially robust object detection," in *ICCV*, 2019.

[20] S Ma, L Sigal, and S Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *CVPR*, 2016.

[21] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han, "Weakly supervised action localization by sparse temporal pooling network," in *CVPR*, 2018.

[22] Z Liu et al., "Weakly supervised temporal action localization through contrast based evaluation networks," in *ICCV*, 2019.

[23] Y et al Yuan, "Marginalized average attentional network for weakly-supervised learning," *arXiv preprint arXiv:1905.08586*, 2019.

[24] Pilhyeon Lee et al., "Background suppression network for weakly-supervised temporal action localization," in *AAAI*, 2020.

[25] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang, "Weakly-supervised action localization by generative attention modeling," in *CVPR*, 2020.

[26] Huijuan Xu, Abir Das, and Kate Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *ICCV*, 2017.

[27] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin, "Temporal action detection with structured segment networks," in *CVPR*, 2017.

[28] AJ Piergiovanni and Michael Ryoo, "Temporal gaussian mixture layer for videos," in *International Conference on Machine learning*. PMLR, 2019.

[29] AJ Piergiovanni and Michael S Ryoo, "Learning latent super-events to detect multiple activities in videos," in *CVPR*, 2018.

[30] Mavroudi, "Representation learning on visual-symbolic graphs for video understanding," in *ECCV*, 2020.