Relaxed Transformer Decoders for Direct Action Proposal Generation **Supplementary Material**

Jing Tan* Jiaqi Tang* Limin Wang✉ Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, China
{jtan,jqtang}@smail.nju.edu.cn, {lmwang, gswu}@nju.edu.cn

Appendices

A. Additional Ablation Studies

A.1. Boundary Attentive Module

Study on attentive representations. To further analyze the design of the boundary-attentive module, we perform ablations on projection placement (i.e., location of MLP) and boundary enhancement methods. For projection placement, we consider the alternatives of pre-enhancement and post-enhancement. For boundary enhancement methods, we also tried concatenating boundary weights with features along the channel dimension. Results of Table A show that pre-enhancement projection has a better performance and multiplication enhancement outperforms concatenation enhancement. We analyze that pre-enhancement projection provides a more compact representation for boundary enhancement and attention introduces a more direct and explicit feature enhancement strategy.

Table A. Ablation study on MLP encoders on THUMOS14, measured by AR@AN.

Projection placement	Boundary enhancement	@50	@100	@200	@500
Pre	multiply	**41.52**	**49.32**	**56.41**	**62.91**
Post	multiply	38.81	47.36	54.86	62.30
Pre	concat	37.92	45.33	52.12	60.86

Study on temporal positional embedding in encoder. In this section, we show the importance of temporal positional embedding in the boundary attentive module. We experiment with removing positional embedding at MLP encoder or directly adding it into encoder. We contend that concatenating positional embedding with video features explicitly gives the encoded features the relative order of the sequence, and simplifies the difficulty of proposal generation by having temporal locations encoded in the features. The results in Table B show that the model performance decreases by 4.4% on AR@50, without temporal positional embedding in encoder.

Table B. Ablation study on position embedding of MLP encoder on THUMOS14, measured by AR@AN.

Positional embedding in encoder	@50	@100	@200	@500
w/o	37.07	45.05	51.58	58.31
w/	**41.52**	**49.32**	**56.41**	**62.91**

Effect of feature receptive field on MLP encoder. Table C is an extension of Table 4 in Section 4.3 to prove that the over-smoothing effect of encoder self-attention causes performance drop. We extend our experiment to alleviate the possibility that features with smaller receptive field boosts the performance in general. By comparing MLP encoder performance of input features with receptive field size of 16 and 64, we conclude that smaller receptive field would decrease the performance of MLP encoder. The increase of performance with Transformer encoder is because that smaller receptive field reduces the over-smoothing effect for Transformer encoder.

Table C. Ablation study on the effect of feature receptive field on MLP encoder on THUMOS14, measured by AR@AN.

Size of Receptive field	@50	@100	@200	@500
64	**41.52**	**49.32**	**56.41**	**62.91**
16	39.56	47.36	53.82	60.47

A.2. Relaxed Transformer Decoder

Study on the relaxation mechanism. We present the two-step **"top-1 to top-k"** matching scheme. In our strategy, we first train with the strict bipartite matching criteria to generate sparse predictions, then fine-tune with the relaxed matching scheme to improve the overall recall. The first step of our strategy is *necessary* because it makes the positive samples sparsely distributed and minor-overlapped, thus the model is free of NMS.

*: Equal contribution. ✉: Corresponding author.

In the fine-tuning phase, we freeze the modules except for binary classification and boundary embeddings. Specifically, we calculate tIoU between targets and predictions, and employ three different settings of the relaxation mechanism. *First*, we mark predictions with tIoU higher than a threshold as positive samples and get an updated matching permutation σ'. We calculate both classification and localization loss according to the updated assignment σ'. *Second*, only loss for the binary classification head is calculated with σ'. The target of this relaxation setting is to improve the quality measurement (confidence) of positive (but not optimal) proposals, and stabilize the distribution of optimal predictions. The *last* one is assigning the closest prediction of each groundtruth as positive elements (predictions of bipartite matching are not necessarily the geometrically closest), and calculate losses on this updated assignment σ''. As Table D illustrates, the results of all three settings are close, demonstrating the influence of the relaxation mechanism is robust to settings (rule and scope).

With the relaxation mechanism, our model witnesses an evident improvement on AR and AUC. With the optimal bipartite matching, RTD-Net predicts proposals of bipartite matching (top-1 proposals) well, while it suppresses several other predictions around the groundtruth (top-k proposals), which results in a decrease of AR at large AN and overall AUC. In the fine-tuning phase, our model improves the scoring of top-k proposals with the relaxation mechanism, and the performance of top-1 proposals is not affected. As a result, the relaxation mechanism boosts the overall performance of RTD-Net.

Similar to us, [10] exploits a "stop-grad" operation, namely they freeze the FCOS detector [6] and train their PSS head in the fine-tuning phase. The difference is that [10] firstly makes top-k predictions well and then learns to predict top-1 proposals. RTD-Net exploits a **"top-1 to top-k"** strategy, while [10] leverages a **"top-k to top-1"** scheme. Both of them aim to optimize the procedure of label assignment at the cost of removing heuristic NMS, and markedly reduce the inference time.

Table D. Ablation study on the rule of relaxation mechanism on ActivityNet-1.3 validation set, measured by AR@AN and AUC.

Rule	Scope	AR@1	AR@100	AUC
None	None	32.73	71.88	65.50
threshold	cls + loc	33.05	73.21	**65.78**
threshold	cls	**33.10**	73.12	65.77
top1	cls + loc	32.95	**73.25**	65.77

Study on temporal positional embedding in decoder. Explicit temporal positional embedding also plays a key role in the relaxed transformer decoder. We experiment with no positional embedding, add positional embedding at encoder-decoder attention input and similar to detr, add positional embedding only at attention. As shown in Table E, adding positional embedding at attention achieves the best performance. RTD-Net achieves 37.43% on AR@50 without positional embedding in the decoder, which decreases by about 4%. Adding positional embedding at input causes performance drop as well, by 2.0% on AR@50.

Table E. Ablation study on position embedding of transformer decoder on THUMOS14, measured by AR@AN.

Positional embedding	@50	@100	@200	@500
None	37.43	46.01	53.90	61.32
At input	39.53	47.13	53.83	61.67
At attn.	**41.52**	**49.32**	**56.41**	**62.91**

Study on the number of decoder layers. We conduct experiments on the number of decoder layers and the results are displayed in Table F. RTD-Net achieves the best performance with 6 decoder layers, in terms of AR@AN. When the number of decoder layers increases from 1 to 2, it improves AR@50 by around 6.2, but this improvement decreases to 1.8 when the number of decoder layers increases from 2 to 3.

Table F. Ablation study on the number of decoder layers on THUMOS14, measured by AR@AN.

Number of decoder layers	@50	@100	@200	@500
1	32.76	42.93	51.09	58.19
2	38.92	47.47	53.14	60.11
3	40.71	47.57	53.84	60.30
6	**41.52**	**49.32**	**56.41**	**62.91**
9	38.36	46.70	53.70	60.01

A.3. Non-Maximum Suppression

In Table G, we conduct experiments on RTD-Net with and without NMS, and observe similar results. NMS is not necessary in RTD-Net because the predictions are relatively sparse and minor-overlapped with our two-step training strategy (details in Appendices A.2). In contrast, BSN [5] and BMN [4] generate highly overlapped proposals with similar confidence, as shown in Figure G of Appendices. Therefore, NMS is needed for these dense proposal generators to suppress such proposals.

Table G. Ablation study on non-maximum suppression on THUMOS14, measured by AR@AN.

Method	@50	@100	@200	@500
RTD-Net	41.52	49.32	**56.41**	**62.91**
RTD-Net+SNMS	**42.02**	**49.40**	54.98	61.16

B. Visualization

Visualization of boundary-attentive representations. Figure A(a) shows the pattern for input video feature. Vertical line patterns are visible in input features, indicating

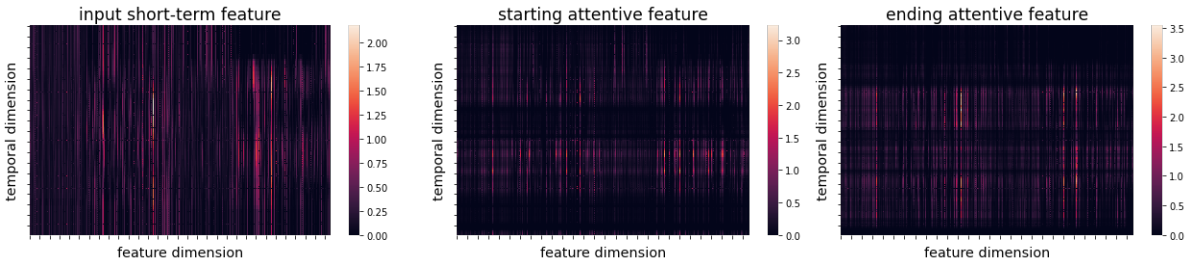

Figure A. (a) is visualization of input short-term feature of a randomly sampled video segment, this feature has a receptive field of 64 frames; (b) is visualization of starting and ending attentive features. Best viewed in color.

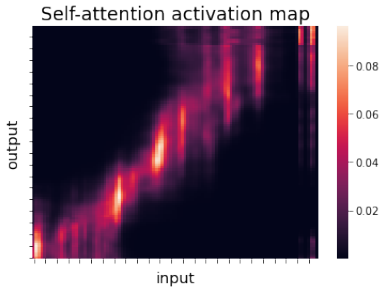

Figure B. Visualization of self-attention activation map in Transformer encoder. Best viewed in color.

different temporal locations sharing similar feature representation. That is the slowness phenomena that we discover in short-term video features. To alleviate this slowness, we explicitly multiply starting and ending attentive scores with features. Figure A(b) illustrates the starting and ending attentive feature. We observe the aforementioned vertical line patterns are broken by horizontal darker line patterns, indicating that effectiveness of boundary information in representation enhancement.

Analysis on the over-smoothing effect. We further explore the reason for the over-smoothing effect with self-attention mechanism of the transformer encoder. Figure B shows the self-attention map of a sample from THUMOS14 [3]. The x-axis is the input temporal locations, and the y-axis is the output temporal locations. A diagonal activation pattern is observed in Figure B, with many short vertical line patterns visible around the diagonal activation. The vertical patterns indicate that many different output locations share the same input activation, which result in the over-smoothing effect. The input short-term feature already has the problem of slowness, adding temporal attention to this feature would aggravate the slowness and result in weaker performance.

Visualization of decoder attention maps. In this subsection, we present the activation map from self-attention layer and encoder-decoder attention layer in RTD decoder layers. Figure C shows the $N_Q \times N_T$ (N_T is the number of time steps in each snippet, N_Q is the number of queries predicted for each snippet) encoder-decoder activation map from Layer 1, 3 and 5 (last) of decoder layers from a randomly selected video snippet. Vertical patterns are visible in these activation maps. Each blue vertical beam corresponds to the ending of an action instance, which indicates that proposal queries are more focused on the features from the ending region of an action.

Figure D shows the $N_Q \times N_Q$ query self-attention activation map from the last layer of decoder. High activations are visible along the y-axis, indicating that proposal queries are keen at learning from some well predicted queries (eg. 1st, 14th and 27th) at inter-proposal modeling. The 14th query in Figure D is the highest ranked and also a well-predicted proposal in results.

C. Additional Comparisons with SOTA

AR curves under all tIoU thresholds. RTD-Net generates more precise and more complete proposals, compared with previous methods. We compare RTD-Net with bottom-up method BSN under different tIoU thresholds for recall. In Figure E, we demonstrate that: 1) RTD-Net outperforms BSN under every tIoU threshold, especially at smaller number of proposal conditions. 2) RTD-Net outperforms BSN under high tIoU thresholds, indicating that when the true positive standard is strict with localization, RTD-Net still achieves higher recall with better localized predictions.

Efficiency Analysis. Our RTD-Net only presents the transformer decoder, while keeping the original MLP encoder for feature extraction. Therefore, our encoder is with linear run-time and memory complexity. Our decoder uses cross attention and the complexity is $O(N_T \times N_Q)$. In practice, N_Q could be smaller than sequence length. In our experiment, we found our method uses 1,519 MB GPU memory while existing SOTA methods such as BMN uses 7,152 MB. In addition, we provide a run-time breakdown for RTD-Net and BSN in Table H. We infer with 3-minute video input on one RTX 2080-Ti GPU. We follow [5, 4] to exclude the backbone feature extractor. It is noted that, for a 3-minute video, RTD-Net predicts 640 proposals without any post-processing module while BSN outputs about 3k predictions for the time-consuming SNMS post-processing.

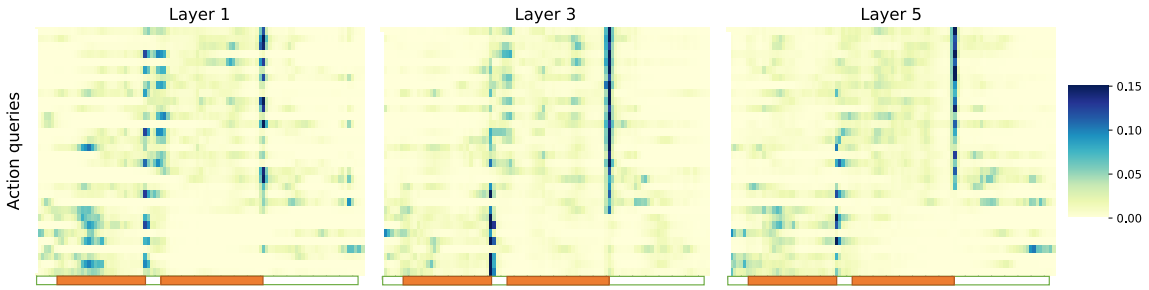

Figure C. Visualization of encoder-decoder attention activation map, averaged among multiple heads. The y-axis is action queries and the x-axis represents time steps from encoder features. From yellow to blue represents the intensity of activation, the bluer the stronger the activation. The white and orange bar underneath the x-axis demonstrates groundtruth instances in this snippet. The orange part represents action and the rest represents background. Best viewed in color.

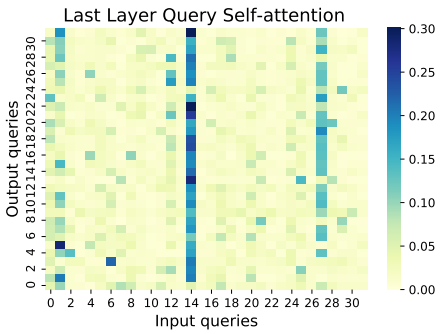

Figure D. Visualization of the self-attention layer in the last layer of Transformer decoder, averaged among multiple heads. Best viewed in color.

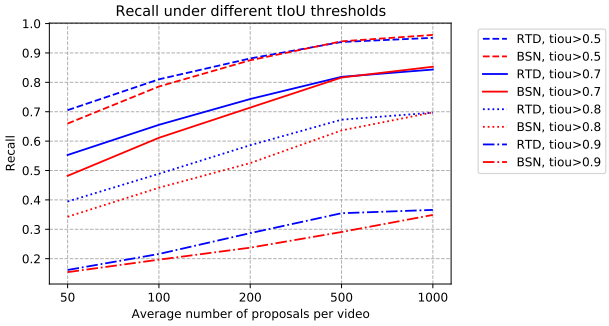

Figure E. Visualization of Average Recall at different proposal numbers under all tIoU thresholds. Best viewed in color.

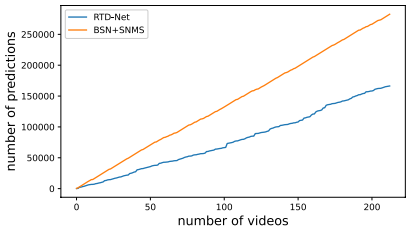

Figure F. Comparison of number of proposals between RTD-Net and BSN.

Table H. Run-time breakdown analysis of RTD-Net and BSN.

(a) RTD-Net

	Boundary-probability predictor + re-weight	MLP encoder	Transformer decoder	Three-branch Head
RTD-Net	49.29ms	0.32ms	8.97ms	0.89ms

(b) BSN

	TEM	PGM	PEM	SNMS
BSN	53.23ms	243.79ms	7.68ms	6026.34ms

RTD-Net directly generates high-quality proposals with a smaller number of predictions. Due to the pair-wise modeling in our decoder, our predictions do not suffer from the flooding of redundant, highly-overlapping proposals. As shown in Figure F, RTD-Net predicts fewer proposals than BSN [5], but still achieves higher average recall under all metrics on THUMOS14.

Generalizability of proposals. The ability of generating high quality proposals for unseen action categories is an important property of a temporal action proposal generation method. Following BSN [5] and BMN [4], we choose two non-overlapped action subsets: "Sports, Exercise, and Recreation" and "Socializing, Relaxing, and Leisure" of ActivityNet-1.3, as *seen* and *unseen* subsets separately. *Seen* subset contains 87 action classes with 4455 training and 2198 validation videos, and *unseen* subset contains 38 action classes with 1903 training and 896 validation videos. Based on I3D features, we train RTD-Net with *seen* and *seen+unseen* training videos separately, and evaluate on both *seen* and *unseen* validation videos. Results in Table I demonstrate that the performance remains competitive in unseen categories, suggesting that RTD-Net achieves great generalizability to generate high quality proposals for unseen classes, and is able to predict accurate temporal action proposals regardless of semantics.

Qualitative results. We visualize qualitative results in Figure G. The top-5 predictions of BMN [4] share similar starting seconds and scores, and the same ending seconds. Bottom-up methods like BMN retrieve all proposals around locations with high boundary scores, while many of them

Table I. Generalization evaluation of RTD-Net on ActivityNet-1.3.

	Seen(val)		*Unseen*(val)	
	AR@100	AUC	AR@100	AUC
Seen+*Unseen*(train)	70.25	62.66	73.09	65.52
Seen(train)	69.80	61.32	**72.27**	**64.54**

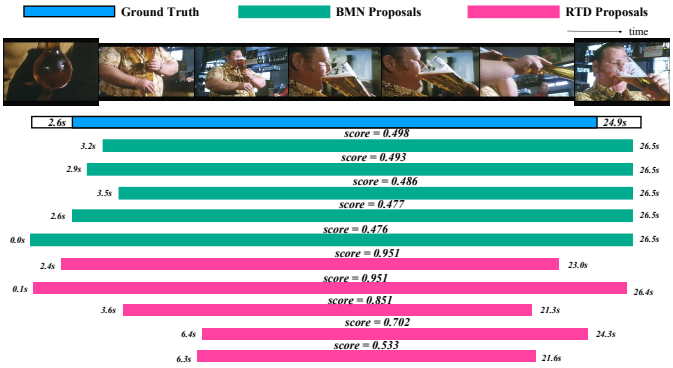

Figure G. Qualitative results of RTD-Net on ActivityNet-1.3. The proposals shown are the top-5 predictions for corresponding groundtruths based on the scoring scheme for each model.

are redundant and evaluated with similar confidence. If proposals around another groundtruth all have confidence over 0.9, the rankings of these proposals with confidence around 0.5 fall down, resulting in a low recall of this groundtruth. Therefore, heuristic NMS is introduced to address the above issues, which increases the inference time drastically. In contrast, a variation in localization appears in RTD predictions. Starting and ending locations of RTD proposals are varying from one another. More importantly, scores of RTD proposals are consistent with their rankings. Incomplete predictions are evaluated with lower scores, and ranked after those well-predicted proposals. As a result, RTD-Net is free of NMS module and has a much faster inference speed.

D. Performance on HACS Segments

Dataset. HACS Segments dataset [8] contain 50,000 untrimmed videos and share the same 200 action categories with ActivityNet-1.3 dataset [2]. To evaluate the quality of proposals, we calculate Average Recall with Average Number of proposals per video (AR@AN), and the Area under the AR vs AN curve (AUC) as metrics on HACS Segments dataset, which are the same as ActivityNet-1.3 dataset.

Comparison with state-of-the-art methods. We simply train RTD-Net on HACS Segments, with the same settings on ActivityNet-1.3. As Table J illustrates, RTD-Net achieves comparable results with only 100 queries per video. In contrast, BSN [5] predicts a large number of proposals and calculates evaluation metrics with top-100 of them. With top-100 proposals, BSN achieves a higher AR@100 than RTD-Net, while AUC of BSN and RTD-Net is the same. The comparison demonstrate RTD-Net achieves higher AR at small AN (e.g., AR@1), which indicates the efficiency of the direct action proposal generation mechanism.

Table J. Comparison with other state-of-the-art proposal generation methods on validation set of HACS Segments in terms of AR@AN and AUC. Among them, only RTD-Net is free of NMS.

Method	TAG+NMS [9]	BSN+SNMS [5]	RTD-Net
AR@1 (val)	-	-	**16.34**
AR@100 (val)	55.88	**63.62**	61.11
AUC (val)	49.15	**53.41**	**53.41**

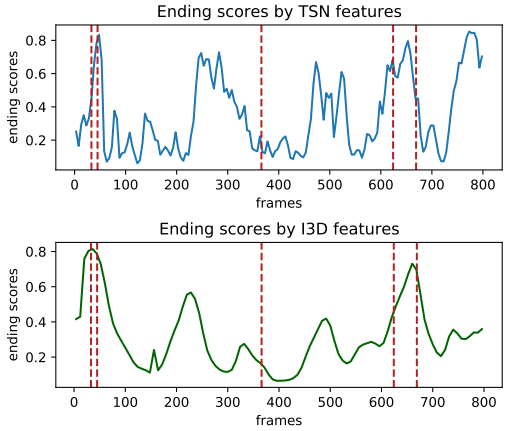

Figure H. Comparison of ending scores predicted by TSN and I3D feature extractors.

E. Feature Encoding

Choices of feature extractors. There are two main types of feature extractors, one is 2D CNN (e.g., TSN [7]), the other captures temporal relations (e.g., I3D [1]). Bottom-up methods (e.g., BSN and BMN) first evaluate boundary confidence of all locations, and then explicitly match starting and ending points. With 2D CNN features that preserve local information better, bottom-up methods can achieve a higher recall of boundaries and better performance, which can be proved in the next section. Compared with 2D CNN features, I3D features have larger receptive fields and contain more temporal contexts. RTD-Net exploits self-attention blocks for proposal-proposal relations, and leverages encoder-decoder blocks to learn action-background differences. Therefore it can make full use of contextual information of I3D features and directly generate center locations and duration of proposals.

Comparison of boundary scores on different feature extractors. According to the mechanism of the temporal evaluation module, temporal locations with boundary scores higher than a threshold or being with peak scores (namely their boundary scores S_i are higher than their neighbors

S_{i-1} and S_{i+1}) are considered as candidates of action boundaries. Figure H displays the ending scores by TSN and I3D features, and groundtruth ending points are marked with vertical red dotted lines. We observe that TSN predictions covers every groundtruths with its local maximas but the first, achieving high recall of ending prediction. In contrast, the temporal evaluation module based on I3D features only captures the first groundtruth, resulting in a weaker recall. This might explains the performance drop of BSN and BMN with I3D feature input and gives solid support for our feature choice of the temporal evaluation module.

Effect of feature modality. In Table K, we show the effect of feature modality on our framework by comparing the performance of RTD-Net under features from different modalities. We experiment with features from RGB, Optical flow and the fusion of both modalities. We find that Flow features outperforms RGB features by 1.5% on AR@50, which indicates that motion information is more significant than appearance information in temporal action proposal generation. The fusion of both modalities here are in an early-fusion fashion, which requires both features concatenated in the beginning of the training and inference of the network. The early fusion features outperforms Flow features by 2.7% on AR@50.

Table K. Comparison of RGB and optical flow on THUMOS14, measured by AR@AN.

Modality	@50	@100	@200	@500
RGB	37.28	45.49	52.73	60.61
Flow	38.75	47.30	54.11	61.11
Early Fusion	**41.52**	**49.32**	**56.41**	**62.91**

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 5

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 5

[3] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 3

[4] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3888–3897, 2019. 2, 3, 4

[5] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–21, 2018. 2, 3, 4, 5

[6] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *ICCV*, pages 9626–9635. IEEE, 2019. 2

[7] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 5

[8] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. HACS: human action clips and segments dataset for recognition and temporal localization. In *ICCV*, pages 8667–8677. IEEE, 2019. 5

[9] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2933–2942, 2017. 5

[10] Qiang Zhou, Chaohui Yu, Chunhua Shen, Zhibin Wang, and Hao Li. Object detection made simpler by eliminating heuristic NMS. *CoRR*, abs/2101.11782, 2021. 2