Dual DETRs for Multi-Label Temporal Action Detection

Supplementary Material

A. More Details

Detection Head. Following previous studies [3, 21, 22, 33], we apply a linear projection to the instance-level content vectors $\boldsymbol{i}^{con}$ to generate the classification score:

$$\hat{\boldsymbol{p}} = \text{Linear}(\boldsymbol{i}^{con}). \tag{11}$$

The classification score, $\hat{\boldsymbol{p}}$, will be used in three scenarios: 1) selecting encoder proposals in the query alignment strategy, 2) performing bipartite matching for assigning ground truth, and 3) calculating the classification loss. Moreover, we employ a Multi-Layer Perception (MLP) with ReLU activation to generate proposal offsets. Specifically, the boundary-level content vectors $\boldsymbol{s}^{con}, \boldsymbol{e}^{con}$ are used to compute boundary-level offsets, while the instance-level content vectors $\boldsymbol{i}^{con}$ are utilized to generate instance-level offsets:

$$\begin{aligned} \Delta \boldsymbol{s} &= \text{MLP}(\boldsymbol{s}^{con}), \\ \Delta \boldsymbol{e} &= \text{MLP}(\boldsymbol{e}^{con}), \\ \Delta \boldsymbol{i} &= \text{MLP}(\boldsymbol{i}^{con}). \end{aligned} \tag{12}$$

These offsets are subsequently employed to refine their respective position vectors.

The detection head is appended to the final encoder layer as well as each decoder layer. Detection losses are computed at these stages to optimize the model. Furthermore, to preserve the alignment between dual-level queries, we share ground truth obtained through bipartite matching among each aligned query.

Training Details. DualDETR is trained on two NVIDIA TITAN Xp GPUs, with a batch size of 16 per GPU. To ensure stable training, we employ ModelEMA [9] and gradient clipping following [29]. The random seed is fixed at 42 to ensure reproducibility.

B. Additional Experiments

Traditional TAD Benchmarks. The performance of DualDETR on traditional benchmarks, THUMOS14 [10] and AcitvityNet1.3 [2], is presented in Tab. 6. DualDETR surpasses previous query-based methods by a significant margin at all IoU thresholds on THUMOS14, achieving an impressive average mAP gain of $\mathbf{10.1}\%$. When compared to standard methods that rely on NMS post-processing, DualDETR exhibits comparable performance to the state-of-the-art method ActionFormer [29]. Moreover, on ActivityNet1.3, DualDETR also outperforms all previous query-based methods. These results further demonstrate DualDETR's superiority in action detection tasks.

Study on Number of Queries. In Tab. 7, we analyze the effectiveness of the number of decoder queries. We find that the optimal number of queries is 150, 25, and 96 for MultiTHUMOS, Charades, and TSU, respectively. This observation aligns with the number of ground truth instances per video in each dataset, which is approximately 97, 6.8, and 77 for MultiTHUMOS, Charades, and TSU, respectively.

Study on Number of Layers. In Tab. 8, we examine the impact of the number of encoder and decoder layers on the MultiTHUMOS dataset. Our default configuration includes 6 encoder layers and 5 decoder layers. Thanks to the joint initialization strategy, the performance remains consistently strong even with a reduced number of decoder layers. In terms of average performance, our default setting proves to be the most effective.

Inference Efficiency. We report the efficiency comparison on multiTHUMOS with two competitive methods ActionFormer [29] and TriDet [19] in Tab. 9. DualDETR achieves the highest mAP with the least latency among all methods, perfectly balancing the efficiency-performance trade-off. The suffix of DualDETR in the table indicates the number of decoder queries employed.

Qualitative Results. To further compare different detection paradigms, we present qualitative results in Fig. 6. Boundary-level detection demonstrates high accuracy in boundary detection but lacks reliable semantic labels. On the other hand, instance-level detection achieves robust detection but sub-optimal boundary localization. Our proposed DualDETR combines both paradigms effectively, offering reliable recognition and precise boundary localization simultaneously. Additionally, we provide qualitative results for high-overlap action regions in Fig. 7. Our method excels in handling complex situations, showcasing the strong applicability of DualDETR in multi-label action detection scenarios.

C. Limitation and Future Work

In the query alignment strategy, where each query is matched with an encoder proposal, the maximum number of queries is constrained by the number of features in the encoder feature map. If situations arise where a larger number of queries is required, additional modules must be devised. Furthermore, to maintain efficiency during training and testing, DualDETR operates on pre-extracted video features following previous practice, which overlooks the gap between pre-training and downstream tasks. However, with the emergence of parameter-efficient fine-tuning techniques like LoRA [8], Adapter [7], and Prompt Tuning [25, 26]

Method	Backbone	THUMOS14 [10]						ActivityNet-v1.3 [2]			
		0.3	**0.4**	**0.5**	**0.6**	**0.7**	**Avg.**	**0.5**	**0.75**	**0.95**	**Avg.**
Standard Methods											
BMN [13]	TSN [23]	56.0	47.4	38.8	29.7	20.5	38.5	50.1	34.8	8.3	33.9
G-TAD [27]	TSN [23]	54.5	47.6	40.2	30.8	23.4	39.3	50.4	34.6	9.0	34.1
BC-GNN [1]	TSN [23]	57.1	49.1	40.4	31.2	23.1	40.2	50.6	34.8	9.4	34.3
TAL-MR [32]	I3D [4]	53.9	50.7	45.4	38.0	28.5	43.3	43.5	33.9	9.2	30.2
TCA-Net [17]	TSN [23]	60.6	53.2	44.6	36.8	26.7	44.3	52.3	36.7	6.9	35.5
BMN-CSA [1]	TSN [23]	64.4	58.0	49.2	38.2	27.8	47.7	52.4	36.2	5.2	35.4
VSGN [31]	TSN [23]	66.7	60.4	52.4	41.0	30.4	50.2	52.4	36.0	8.4	35.1
ContextLoc [34]	I3D [4]	68.3	63.8	54.3	41.8	26.2	50.9	56.0	35.2	3.6	34.2
RCL [24]	I3D [4]	70.1	62.3	52.9	42.7	30.7	51.0	51.7	35.3	8.0	34.4
AFSD [12]	I3D [4]	67.3	62.4	55.5	43.7	31.1	52.0	52.4	35.3	6.5	34.4
DCAN [5]	TSN [23]	68.2	62.7	54.1	43.9	32.6	52.3	51.8	36.0	9.5	35.4
TAGS [16]	I3D [4]	68.6	63.8	57.0	46.3	31.8	52.8	56.3	**36.8**	**9.6**	**36.5**
MUSES [14]	I3D [4]	68.9	64.0	56.9	46.3	31.0	53.4	50.0	35.0	6.6	34.0
Zhu et al. [35]	I3D [4]	72.1	65.9	57.0	44.2	28.5	53.5	**58.1**	36.3	6.2	35.2
ActionFormer [29]	I3D [4]	82.1	77.8	71.0	59.4	43.9	66.8	53.5	36.2	8.2	35.6
TriDet [19]	I3D [4]	**83.6**	**80.1**	**72.9**	**62.4**	**47.4**	**69.3**	–	–	–	–
Query-Based Methods											
TadTR [15]	I3D [4]	62.4	57.4	49.2	37.8	26.3	46.6	49.1	32.6	8.5	32.3
RTD-Net [21]	I3D [4]	68.3	62.3	51.9	38.8	23.7	49.0	47.2	30.7	**8.6**	30.8
DINO [30]	I3D [4]	69.8	63.1	53.7	41.5	26.4	50.9	–	–	–	–
ReAct [18]	TSN [23]	69.2	65.0	57.1	47.8	35.6	55.0	49.6	33.0	8.6	32.6
Self-DETR [11]	I3D [4]	74.6	69.5	60.0	47.6	31.8	56.7	52.3	33.7	8.4	33.8
DualDETR	I3D [4]	**82.9**	**78.0**	**70.4**	**58.5**	**44.4**	**66.8**	**52.6**	**35.0**	7.8	**34.3**

Table 6. **DualDTER's performance on THUMOS14 and ActivityNet1.3.** The results of other methods are mainly from Self-DETR [11].

# Queries (N_q)	80	120	150	180	250
MultiTHUMOS [28]	32.33	32.50	**32.64**	32.60	32.63
# Queries (N_q)	10	25	40	55	70
Charades [20]	14.55	**15.62**	15.27	14.26	13.58
# Queries (N_q)	20	40	60	80	96
TSU [6]	18.56	20.19	20.59	20.73	**20.81**

Table 7. **Ablation study on the number of decoder queries.**

L_E	L_D	0.1	0.3	0.5	0.7	0.9	Avg.
5	5	51.90	45.52	33.54	19.02	3.83	31.25
6	3	52.69	46.66	34.52	19.47	3.75	31.93
6	4	53.39	**47.42**	**35.19**	19.93	3.88	32.52
6	5	**53.42**	47.41	35.18	**20.18**	4.02	**32.64**
6	6	52.95	46.30	34.06	19.47	**4.30**	31.90

Table 8. **Ablation study on the number of encoder and decoder layers on MultiTHUMOS.**

Method	GPU	Param(M)	GMACs	Latency	mAP
ActionFormer	A100	27.90	45.3	224ms	29.6
TriDet	A100	15.25	43.7	167ms	30.7
DualDETR$_{q80}$	TITAN Xp	21.77	66.3	65ms	32.3
DualDETR$_{q150}$	TITAN Xp	21.77	80.3	69ms	32.6

Table 9. **Inference efficiency.**

there is a growing opportunity to explore efficient end-to-end approaches for action detection tasks.

References

[1] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 121–137. Springer, 2020. 2

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1, 2

Figure 6. **Qualitative comparison** between predictions of single-level detection (boundary-level only and instance-level only) and our dual-level detection. The videos are from the MultiTHUMOS dataset.

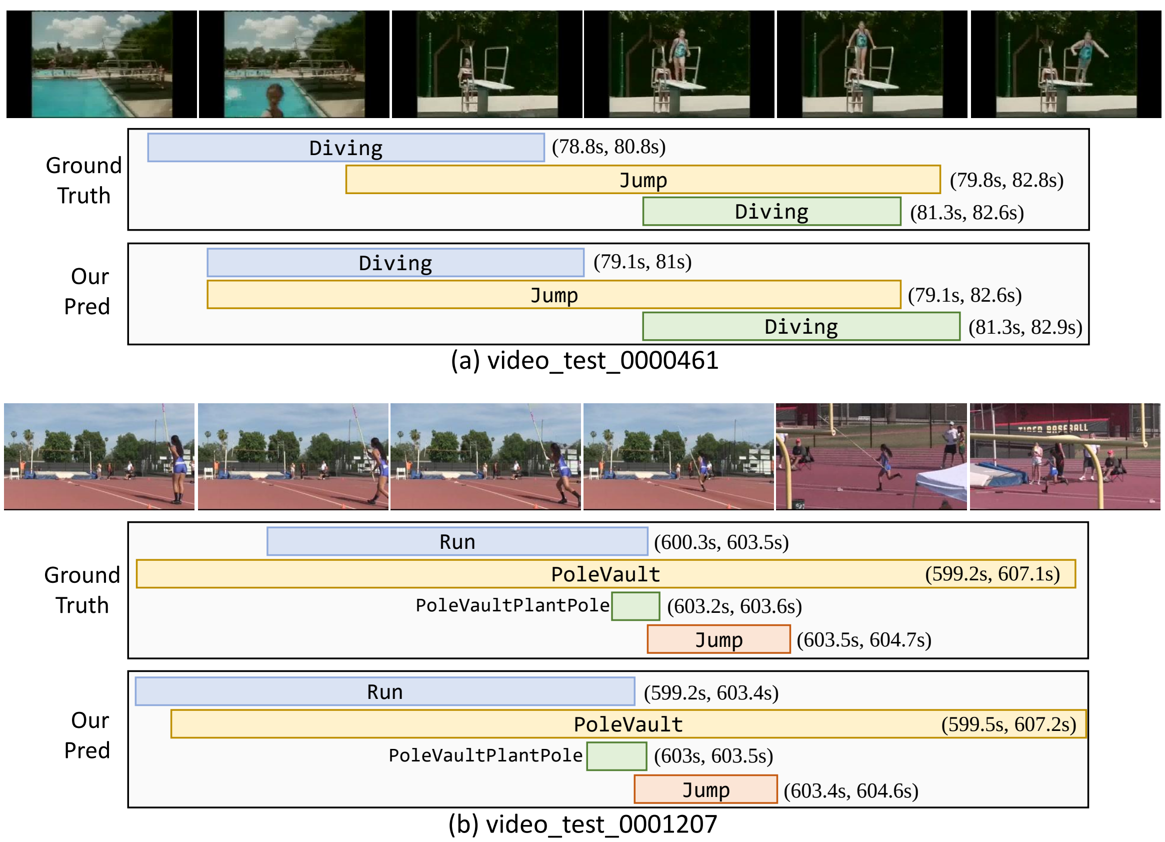

Figure 7. **Qualitative results** under high-overlap scenarios. The videos are from the MultiTHUMOS dataset.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[4] 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. 2

[5] Guo Chen, Yin-Dong Zheng, Limin Wang, and Tong Lu. Dcan: Improving temporal action detection via dual context aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 248–257, 2022. 2

[6] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2533–2550, 2022. 2

[7] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 1

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[9] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017. 1

[10] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 1, 2

[11] Jihwan Kim, Miso Lee, and Jae-Pil Heo. Self-feedback detr for temporal action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10286–10296, 2023. 2

[12] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2021. 2

[13] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 2

[14] Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip HS Torr. Multi-shot temporal event localization: a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12596–12606, 2021. 2

[15] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022. 2

[16] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Proposal-free temporal action detection via global segmentation mask learning. In *European Conference on Computer Vision*, pages 645–662. Springer, 2022. 2

[17] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 485–494, 2021. 2

[18] Dingfeng Shi, Yujie Zhong, Qiong Cao, Jing Zhang, Lin Ma, Jia Li, and Dacheng Tao. React: Temporal action detection with relational queries. In *Computer Vision–ECCV 2022: 17th European Conference*, pages 105–121. Springer, 2022. 2

[19] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 1, 2

[20] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference,*, pages 510–526. Springer, 2016. 2

[21] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13526–13535, 2021. 1, 2

[22] Jing Tan, Xiaotong Zhao, Xintian Shi, Bing Kang, and Limin Wang. Pointtad: Multi-label temporal action detection with learnable query points. In *NeurIPS*, 2022. 1

[23] 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 *European conference on computer vision*, pages 20–36. Springer, 2016. 2

[24] Qiang Wang, Yanhao Zhang, Yun Zheng, and Pan Pan. Rcl: Recurrent continuous localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13566–13575, 2022. 2

[25] Chen Xu, Haocheng Shen, Fengyuan Shi, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *arXiv preprint arXiv:2304.08386*, 2023. 1

[26] Chen Xu, Yuhan Zhu, Guozhen Zhang, Haocheng Shen, Yixuan Liao, Xiaoxin Chen, Gangshan Wu, and Limin Wang. Dpl: Decoupled prompt learning for vision-language models. *arXiv preprint arXiv:2308.10061*, 2023. 1

[27] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020. 2

[28] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126:375–389, 2018. 2

[29] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Computer Vision–ECCV 2022: 17th European Conference*, pages 492–510. Springer, 2022. 1, 2

[30] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Harry Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2022. 2

[31] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021. 2

[32] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 539–555. Springer, 2020. 2

[33] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1

[34] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13516–13525, 2021. 2

[35] Zixin Zhu, Le Wang, Wei Tang, Ziyi Liu, Nanning Zheng, and Gang Hua. Learning disentangled classification and localization representations for temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3644–3652, 2022. 2