Self-Feedback DETR for Temporal Action Detection

Supplementary Material

Jihwan Kim Miso Lee Jae-Pil Heo†
Sungkyunkwan University
{damien, dlalth557, jaepilheo}@skku.edu

A1. Additional Details

More Training Details. As mentioned in the paper, $\mathcal{L}_{\text{reg}}$ of the paper in $\mathcal{L}_{\text{DETR}}$ defined in Eq. 8 of the paper consists of two types losses: L1 and Interaction-over-Union (IoU) losses. We use the weights for L1 and IoU losses as 2 and 5, respectively as done in the baselines [2,3] for both benchmarks. Moreover, when we define $-\log \hat{p}_{\hat{j}(i)}(c_i)$ in $\mathcal{L}_{\text{DETR}}$ as $\mathcal{L}_{\text{cls}}$, we use the weight for $\mathcal{L}_{\text{cls}}$ as 2 for both datasets. The initial learning rates are $2 \times 10^{-}4$ and $1 \times 10^{-}4$ for THUMOS14 and ActivityNet.

Further Explanation for Guidance Map. The guidance mechanisms are the same for both the encoder and the decoder. Let us explain the guidance map specifically for the decoder's self-attention with an example in Fig. A1. If the 1st and 2nd decoder queries are similar, they will attend similar encoder tokens, as the elements a_{11}, a_{12}, a_{21} and a_{22} with high values in the cross-attention map A_C^1 in the figure. Then, the elements g_{12} and g_{21} in the guidance map G_D^1, calculated by matrix multiplication of A_C^1 and its transpose, will have high values, indicating high correlation b/w the 1st and 2nd queries. Let us review the *ideal* self-attention: the elements a_{12} and a_{21} in the self-attention map A_D^1 should also have high values if the 1st and 2nd decoder queries are similar. It implies G_D^1 is analogous to *ideal* case of A_D^1 so it can be a guidance to *temporally collapsed* A_D^1.

Motivation behind the Design. Decoder queries do not always attend foreground features exclusively; they often include both foreground and background features simultaneously, as in Fig. A2. Background regions serve two essential purposes for TAD. Firstly, they define the boundaries of the actions through the surrounding background frames. Secondly, background features provide contextual information about the actions. Thus, a guidance map exhibiting a strong correlation between foreground and background encoder features is not only intuitive but also valuable.

From sound self-attention maps in DETR of object detection as in Fig. 1(a) of the paper, we observe two key characteristics: 1) correlation between adjacent tokens, 2) diversity. First of all, based on the further explanation for guidance map in the previous paragraph, the cross-attention map encompasses correlations within encoder features or within decoder queries. This motivates us to make references for self-attention maps using cross-relation. Furthermore, we can introduce diversity to self-attention as our guidance maps ensure high diversity. This is because the diversity of guidance map follows high diversity of the cross-attention map. Mathematically, it is trivial that $\text{rank}A = \text{rank}AA^\top = \text{rank}A^\top A$ for a real matrix A.

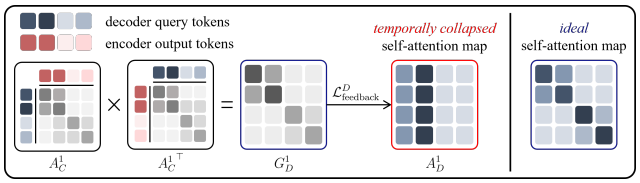

Figure A1: **Further explanation for guidance Map of the decoder.** The figure illustrates an example of constructing guidance map for self-attention of the decoder.

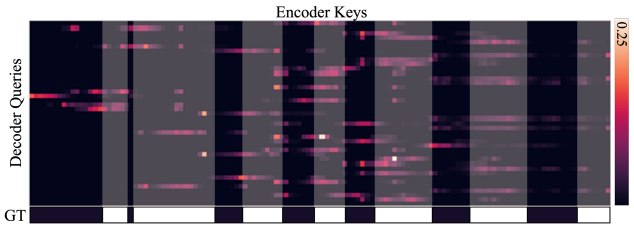

Figure A2: **Visualization of the decoder cross-attention map.** The figure depicts a cross-attention map of the final decoder layer for a validation video in THUMOS14.

A2. Additional Results

Number of Layers. Tab. A1 shows performances with various numbers of layers for the encoder and decoder. Note that we use as default 2 and 4 layers for the encoder and decoder, respectively. As seen in the table, the performance consistently decreases when using a small number of layers

†Corresponding author

Encoder	Decoder	0.3	0.4	0.5	0.6	0.7	Avg.
1	4	74.4	69.2	59.6	45.8	29.2	55.6
2	4	**74.6**	**69.5**	**60.0**	**47.6**	**31.8**	**56.7**
2	3	74.1	67.4	58.6	45.3	29.9	55.1
2	2	67.5	61.9	51.8	39.4	24.9	49.1
2	1	66.0	58.5	48.9	36.2	20.8	46.1

Table A1: **Number of layers for the encoder and decoder.** The table shows performances according to the number of layers for the encoder and decoder.

Encoder	Dec.SA	0.3	0.4	0.5	0.6	0.7	Avg.
·	·	70.1	62.7	51.3	36.5	20.7	48.3
✓	·	**70.7**	**64.7**	**54.0**	38.7	23.3	**50.3**
·	✓	67.8	61.8	52.2	**40.8**	**24.6**	49.4
✓	✓	70.5	64.3	53.9	39.3	23.8	**50.3**

Table A2: **Ablation on collapsed self-attention.** The table shows the results of ablation on collapsed self-attention of DETR without our self-feedback.

Method	0.3	0.4	0.5	0.6	0.7	Avg.
DETR	70.5	64.3	53.9	39.3	23.8	50.3
DETR + Self-feedback	74.5	69.5	60.0	47.6	31.8	56.7
DINO	69.8	63.1	53.7	41.5	26.4	50.9
DINO + Self-feedback	74.7	69.4	59.7	46.8	32.9	56.7

Table A3: **Recent DETR approach on THUMOS14**. The table shows the results on DINO [4], a recent DETR method, with our self-feedback on THUMOS14.

than the default setting.

Ablation on Collapsed Self-Attention. As mentioned in the paper, we argue that the collapsed self-attention modules in the encoder and decoder will play no role for the task. Tab. A2 shows performances of ablation on the collapsed self-attention modules. To ablate self-attention of the encoder, we remove the entire encoder. As for the decoder, we just remove the self-attention modules.

As seen in the table, the performance drop is quite marginal when we ablate the entire encoder or decoder self-attention. From this result, we find that the collapsed self-attention modules hardly help the model to solve TAD.

Recent DETR approach with Self-Feedback. Table. A3 shows the results of deploying a recent DETR approach, DINO [4]. While DINO demonstrates excellent performance in object detection, simply deploying the denoising task does not enhance DETR for TAD. Nevertheless, the self-feedback is still valid for DINO for TAD as temporal collapse persists with DINO.

Self-feedback Losses. We further analyzed the trend of the feedback losses according to the epoch as shown in Fig. A3. Our proposed pipeline helps self-attention maps hold positions in the beginning and helps play their own roles finally through keeping the balance with the main objective.

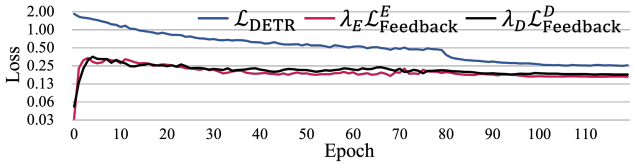

Figure A3: **Losses of main and feedback objectives on THUMOS14.** The figure shows the training losses of main and feedback objectives over epochs on THUMOS14.

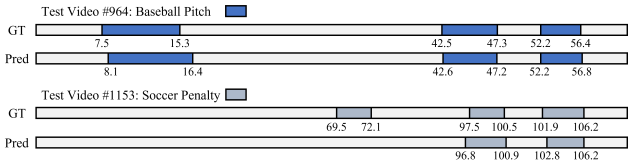

Figure A4: **Qualitative results on THUMOS14.** The figure shows qualitative examples of Self-DETR for two validation videos in THUMOS14.

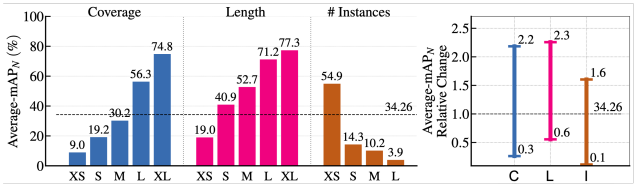

Figure A5: **DETAD analysis on ActivityNet.** It shows the DETAD [1] sensitivity analysis on ActivityNet.

Error Analysis. Fig. A4 illlustrates qualitative results for successful (upper) and failure (lower) cases. Additionally, Fig. A5 depicts sensitivity analysis of DETAD [1] on ActivityNet. In analysis, inferior performance of short scales is a crucial research concern for future work.

References

[1] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *Proceedings of the European conference on computer vision (ECCV)*, pages 256–272, 2018. A2

[2] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *arXiv preprint arXiv:2106.10271*, 2021. A1

[3] 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, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 105–121. Springer, 2022. A1

[4] 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. A2