Prediction-Feedback DETR for Temporal Action Detection

Jihwan Kim, Miso Lee, Cheol-Ho Cho, Jihyun Lee, Jae-Pil Heo[†]

Sungkyunkwan University

{damien, dlalth557, gersys, ibluee01, jaepilheo}@skku.edu

Abstract

Temporal Action Detection (TAD) is fundamental yet challenging for real-world video applications. Leveraging the unique benefits of transformers, various DETR-based approaches have been adopted in TAD. However, it has recently been identified that the attention collapse in self-attention causes the performance degradation of DETR for TAD. Building upon previous research, this paper newly addresses the attention collapse problem in cross-attention within DETR-based TAD methods. Moreover, our findings reveal that cross-attention exhibits patterns distinct from predictions, indicating a short-cut phenomenon. To resolve this, we propose a new framework, Prediction-Feedback DETR (Pred-DETR), which utilizes predictions to restore the collapse and align the cross- and self-attention with predictions. Specifically, we devise novel prediction-feedback objectives using guidance from the relations of the predictions. As a result, Pred-DETR significantly alleviates the collapse and achieves state-of-the-art performance among DETR-based methods on various challenging benchmarks, including THUMOS14, ActivityNet-v1.3, HACS, and FineAction.

Introduction

With the advancement of society, the use of video media has become increasingly widespread. As a result, the demand for efficient methods to search for desired segments within untrimmed videos has grown significantly. One fundamental task, Temporal Action Detection (TAD), aims to identify specific actions within a video and determine their temporal boundaries. TAD has primarily advanced through two-stage approaches. However, recent research has increasingly focused on end-to-end DETR-based methods.

DETR (Carion et al. 2020) is a framework initially proposed and developed in the literature of object detection, introducing the first end-to-end detection framework using set prediction. The DETR method has also been extended to the video domain and applied to TAD (Tan et al. 2021; Liu et al. 2022b; Shi et al. 2022). In TAD, each query is used to predict an action within the video along with its corresponding time interval. To achieve this, bipartite matching is employed to align each query with the ground-truth actions and their temporal intervals within the untrimmed video. This approach has the distinct advantage of eliminating traditional heuristics like Non-Maximum Suppression (NMS).

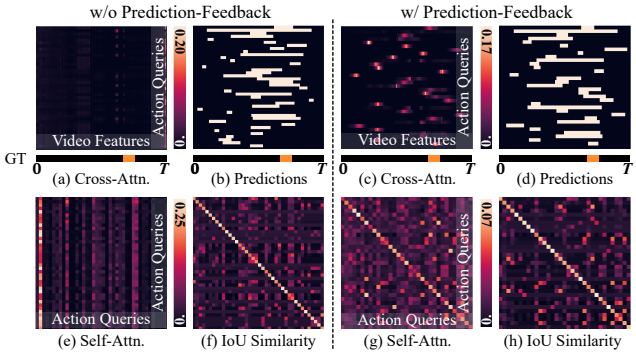

Figure 1: **Attention collapse problem.** The figure depicts the cross- ((a), (c)) and self-attention maps ((e), (g)) of the decoder as well as the predictions ((b), (d)) and their normalized IoU similarity map ((f), (h)). DETR for TAD with standard attention severely suffers from the attention collapse in its cross-attention and self-attention ((a), (e)). The collapsed attention focuses on a few encoder features (a) or decoder queries (e) regardless of the DETR predictions ((b), (f)).

Although DETR with standard attention (shortly original-DETR) has advanced compatibly with Deformable-DETR (Zhu et al. 2021) in object detection, original-DETR in TAD even with recent architectures like DAB-DETR (Liu et al. 2022a) shows a way worse performance. Recently, the root of this issue is identified as the attention collapse problem in self-attention (SA) by Self-DETR (Kim, Lee, and Heo 2023) as depicted in (e) of Fig. 1, where all decoder queries focus on a few queries. The attention collapse is the phenomenon of skipping the attention module to prevent from degeneration of the model (Dong, Cordonnier, and Loukas 2021) towards a rank-1 matrix. Self-DETR utilizes the cross-attention (CA) map to recover the collapsed SA.

However, their solution depends on the soundness of the CA; otherwise it could be suboptimal. We discover that it is not sound; rather collapsed as depicted in Fig. 1. The figure shows that the decoder queries in CA attend to a few encoder features ((a) of the figure), exhibiting the same pattern over almost all queries. It is a particularly critical issue because

[†]Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

CA is crucial for the task as it bridges the queries and the video features. This leads us to resolve the collapse of CA and develop another approach for self-feedback.

Fig. 1 also illustrates the localization predictions in (b), and their corresponding Interaction-over-Union (IoU) map as the self-relation of the queries in (f). In the figure, the attention maps demonstrate clearly different patterns regardless of their predictions and self-relation. Typically, we interpret that the attention maps represent where the model focuses, thus implying why it produces those results. Hence, this phenomenon is analogous to a shortcut, where the model relies on simpler cues rather than learning meaningful representations. Despite this collapsed attention, the model still generates diverse and plausible results, even though all queries focus on the same background regions, as seen in (a) of the CA. This occurs because bipartite matching in the objective of DETR enforces varied predictions by penalizing duplicate results. Based on this observation, we suggest that attention maps be aligned with their corresponding predictions. By using the predictions, rather than the collapsed CA, as a guide for attention, we aim to generalize the model and address the issue of attention collapse.

To this end, we propose a new framework, Prediction-Feedback DETR (Pred-DETR), to tackle the collapse of the entire attention mechanisms in DETR. Our approach begins by expressing the relation of the decoder queries as the IoU similarity map of the DETR predictions with their time intervals. We also reformulate the CA map into the self-relation of the decoder queries. Next, we introduce an auxiliary objective that aligns the self-relation from the CA and SA maps with the IoU similarity map derived from the predictions. Additionally, we leverage encoder predictions from the recent DETR mechanism to guide the encoder SA and decoder CA. Through extensive experiments with various challenging benchmarks including THUMOS14, ActivityNet-v1.3, HACS, and FineAction, we demonstrate that the proposed methods remarkably reduce the degree of the attention collapse problem. Furthermore, the activated attention leads to substantial performance improvements, achieving a new state-of-the-art among DETR-based methods.

To sum up, our main contributions are as follows:

- We identify the attention collapse problem in cross-attention of DETR for TAD. Especially, we found that the cross-attention exhibits clearly different patterns from the predictions, which implies a short-cut phenomenon due to the collapse.
- We propose a novel framework, Prediction-Feedback DETR (Pred-DETR), which utilizes predictions for relieving the attention collapse. We give an auxiliary objective for the collapsed attention modules to be aligned with the IoU relation of the predictions.
- Our extensive experiments demonstrate that Pred-DETR remarkably reduces the degree of the attention collapse by maintaining high diversity of attention. Moreover, we validate that our model achieves a new state-of-the-art performance over the DETR-based models on THUMOS14, ActivityNet-v1.3, HACS, and FineAction.

Related Work

Temporal Action Detection

Temporal action detection (TAD) task focuses on identifying time intervals of action and classifying the instance within untrimmed videos. Over the past decade, significant advancements in TAD have been achieved by foundational methods (Yeung et al. 2016; Shou, Wang, and Chang 2016; Buch et al. 2017). Inspired by the success of two-stage mechanisms in object detection, many TAD methods have adopted a multi-stage framework (Gao et al. 2017; Zhao et al. 2017; Xu, Das, and Saenko 2017; Kim and Heo 2019).

As the subsequent work, point-wise learning has been widely adopted to generate more flexible proposals without pre-defined time windows. SSN (Zhao et al. 2017) and TCN (Dai et al. 2017) introduced extended temporal context around the generated proposals to enhance ranking performance. BSN (Lin et al. 2018) and BMN (Lin et al. 2019) grouped start-end pairs to build diverse action proposals, then scored them for final localization predictions. BSN++ (Su et al. 2021) pointed out the imbalance problem over temporal scales of actions based on BSN. Recently, ActionFormer (Zhang, Wu, and Li 2022) and TriDet (Shi et al. 2023) deployed transformer-based encoder as multi-scale backbone network, and BRN (Kim et al. 2024) resolved the issue of multi-scale features for TAD.

DETR

DETR (Carion et al. 2020) is the first work to view object detection as a direct set prediction problem, and allow for end-to-end detection without any human heuristics such as non-maximum-suppression (NMS). However, DETR demands 10 times longer training than the conventional approaches as bipartite matching is hard to optimize. For this issue, Deformable DETR (Zhu et al. 2021) introduced sparse attention, which attends only a part of elements by learning to specify positions to focus on. The subsequent DETR-based models (Meng et al. 2021; Liu et al. 2022a) further advanced query representations through explicitly encoding box information, which effectively helps to stabilize training.

In TAD, the DETR-based methods are also deployed as DETR has reached a new state-of-the-art performance in object detection. RTD-Net (Tan et al. 2021) identified the problem of the dense attention in the encoder of DETR, which exhibits nearly uniform distribution causing that the self-attention layers act like an over-smoothing effect. TadTR (Liu et al. 2022b) devised temporal deformable attention inspired by Deformable DETR (Zhu et al. 2021). ReAct (Shi et al. 2022) developed a new relation matching to enforce high correlation between queries with low-overlap and high feature similarity. Also, LTP (Kim, Lee, and Heo 2024) proposed a pre-training strategy tailored for DETR.

Recently, Self-DETR (Kim, Lee, and Heo 2023) revealed the problem of the degraded DETR performance for TAD as attention collapse in the self-attention and proposed self-feedback to utilize a guidance map from the cross-attention maps for the self-attention modules. Although it remarkably reduced the degree of the attention collapse, its optimal performance depends on the assumption that cross-attention

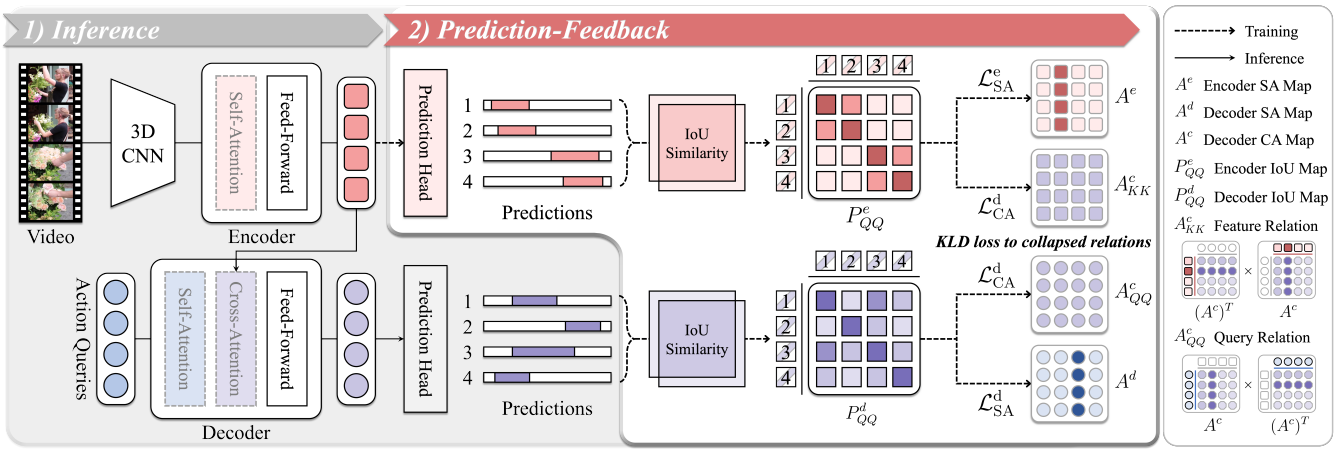

Figure 2: **Overall architecture of the proposed framework, Pred-DETR.** The figure illustrates the entire framework of our model, Pred-DETR. Pred-DETR consists of the two main parts: DETR architecture and prediction-feedback. The encoder and decoder predictions are converted to the relation of Intersection-over-Union (IoU). Then these IoU maps are utilized for prediction-feedback for the collapsed self- and cross-attention. Note that the encoder predictions are deployed only for training.

is sound. However, we discover that cross-attention has collapsed, and therefore introduce prediction-guided feedback which activates the cross-attention as well as the self-attention based on guidance from prediction relations.

Our Approach

This section introduces our proposed method, prediction feedback for Pred-DETR. To be specific, we first elaborate on the preliminaries, and discuss the attention collapse and predictions. Then the explanation of our prediction-feedback mechanisms is followed depicting the overall framework in Fig. 2. Moreover, we provide an extension of prediction-feedback to the encoder via the recent DETR architecture, only for training. Finally, we summarize the overall objectives for Pred-DETR.

Preliminary

DETR. DETR (Carion et al. 2020) adopts the transformer (Vaswani et al. 2017) architecture and composed of two main components: encoder and decoder. First, the encoder captures the global relationships among input features, which is achieved through similarity calculation of SA.

On the other hand, the decoder performs cross-attention operations between object queries and encoder features. Here, object queries are learnable embedding vectors that learn positional information similar to anchors. This mechanism ensures that each query attends to the most relevant parts of the input features processed by the encoder.

Attention Mechanism. An attention module takes three inputs, projecting each into three latent spaces through linear layers. The resulting projections are referred to as query Q, key K and value V, respectively. The attention map is then computed by matrix multiplication of Q with the transpose of K, followed by applying the softmax activation function, scoring similarity between Q and K. By pooling V with the scores followed by a linear projection, we obtain the output of the attention modules. Formally, Q, K, and V are represented as $\mathbb{R}^{N_q \times D}$, $\mathbb{R}^{N_k \times D}$, and $\mathbb{R}^{N_v \times D}$, respectively, where N_q, N_k, and N_v denote the lengths of Q, K and V while D represents the number of channels. When Q, K and V all have the same number of channels, the attention mechanism can be formulated as follows:

$$\text{Attention}(Q, K, V) = AV; A = \text{softmax}(\frac{QK^\top}{\sqrt{D}}), \quad (1)$$

where $A \in \mathbb{R}^{N_q \times N_k}$ is the attention map, $A^\top$ is the transpose of A. For the SA module, the inputs to Q, K and V are the same while Q is obtained from the object queries, while K and V are from the encoder features in the CA module.

DETR for TAD. There are three different things from original DETR for object detection. First of all, we utilize video features from 3D CNN pre-trained on Kinetics (Kay et al. 2017). Note that 3D CNN is frozen and only the temporal dimension is left for video features by global average pooling over the spatial dimensions. Secondly, decoder queries act as action queries instead of object queries since decoder's outputs are used to predict the temporal action detection results. Lastly, DAB-DETR (Liu et al. 2022a) is adopted, consistent with Self-DETR (Kim, Lee, and Heo 2023).

Self-DETR. It is the first work to identify the collapse of encoder and decoder SA maps in DETR when applied to TAD. To guide the collapsed SA maps, they process the CA map $A^c \in \mathbb{R}^{N_q^d \times N_k^d}$ as follows:

$$A^c_{QQ} = A^c \times (A^c)^T, \; A^c_{KK} = (A^c)^T \times A^c, \quad (2)$$

where $A^c_{QQ} \in \mathbb{R}^{N_q^d \times N_q^d}$ and $A^c_{KK} \in \mathbb{R}^{N_k^d \times N_k^d}$ indicate relations between queries and between keys, respectively. In the next step, they ensure the encoder and decoder SA maps resemble A^c_{KK} and A^c_{QQ} by applying Kullback–Leibler (KL) divergence loss. Please refer to the original paper for additional details.

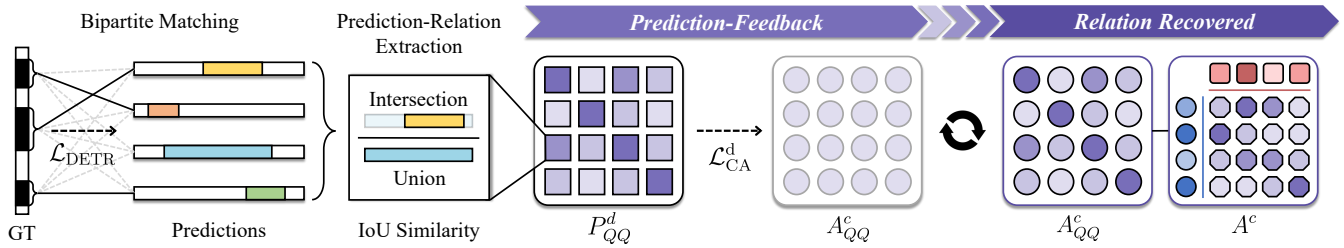

Figure 3: **Prediction-Feedback.** This illustrates the detailed mechanism of prediction-feedback for the cross-attention. The DETR predictions are diverse thanks to the bipartite matching. By aligning attention with the IoU relation from the predictions, the query relation is recovered, alleviating the attention collapse.

Prediction-Feedback

Attention Collapse. The attention collapse problem is a phenomenon where the attention matrix becomes a rank-1 matrix to skip the attention module to prevent degeneration of the learning (Dong, Cordonnier, and Loukas 2021). The collapsed attention outputs uniform values for all queries, resulting in the input being conveyed without additional representations through the residual connection. In this paper, we newly discover the collapse of CA. This problem brings a question for the assumption of the previous work that the CA is sound. As a result, the complete remedy for the collapse through the entire attention modules is required.

Feedback from Predictions. DETR is the first work of end-to-end detection mechanism without anchor boxes or non-maximum suppression (NMS). For this, it uses learnable queries and bipartite matching to assign detection targets to queries since there is no pre-defined matches between predictions and the ground-truth. As the matching is one-to-one mapping, the DETR predictions will be diverse because one query gets a negative loss when two queries produce similar localization results. From this property, the guidance from the predictions can activate the collapse attention modules.

Feedback for Cross-Attention. The bottom line is that both the prediction and the CA map are represented as the relations of the decoder queries to bridge between them. Thereafter, we make the collapsed relation follow the diverse relation from predictions through an auxiliary objective. The CA of the decoder connects the decoder queries with the encoder features to predict actions of interest. One can regularize the collapsed CA map directly; however, to maintain the flexibility of attention results, we propose to guide the query relation A^c_{QQ} extracted from the A^c. The query relation is simply extracted from the CA, indicating how they attend to a similar group of encoder features. Here, the purpose of reformulation of CA into self-relation is the opposite where Self-DETR utilizes A^c_{QQ} as the guidance of feedback.

Subsequently, we design a guidance map based on predictions. The key point here is that the query relation can also be extracted from the IoU similarity among predictions as depicted in Fig. 3. To be specific, each prediction is identical to the refined decoder query and includes the time interval $t_i = \{s_i, e_i\}$ where s_i and e_i are the start and end times, respectively. Therefore, the query relation is obtained from predictions by constructing an IoU similarity matrix $P^d_{QQ} \in \mathbb{R}^{N^d_q \times N^d_q}$, where N^d_q denotes the number of decoder queries, as below.

$$P^d_{QQ}(i,j) = \frac{\max(0, \min(e_i, e_j) - \max(s_i, s_j))}{\max(e_i, e_j) - \min(s_i, s_j)}, \quad (3)$$

where $i, j = 1, 2, \ldots, N^d_q$.

After normalizing P^d_{QQ} by a softmax function, feedback objective with predictions for decoder CA is finally defined as follows:

$$\mathcal{L}^{\text{d}}_{\text{CA}} = D_{KL}(A^c_{QQ} \,||\, P^d_{QQ}), \quad (4)$$

where D_{KL} is the KL divergence loss.

Feedback for Decoder Self-Attention. Furthermore, we propose to guide the collapsed decoder SA maps with P^d_{QQ}. Previous work has already shown the collapsed of the decoder SA maps and the positive impact of feedback on their recovery. In addition, we enhance the feedback mechanism by utilizing P^d_{QQ}, which guarantees higher diversity than CA-based guidance. Prediction-feedback objective for the decoder SA maps $A^d \in \mathbb{R}^{N^d_q \times N^d_q}$ is defined as follows:

$$\mathcal{L}^{\text{d}}_{\text{SA}} = D_{KL}(A^d \,||\, P^d_{QQ}). \quad (5)$$

Feedback for Encoder Self-Attention. Besides the decoder, the encoder SA also suffers from a severe attention collapse. Thanks to the query initialization from the encoder proposed in (Zhu et al. 2021), we obtain the predictions from the encoder. Concretely, we add a linear layer on top of the encoder to predict, utilizing each encoder feature as action queries. This allows us to construct the IoU relation between encoder features just as the IoU relation between decoder queries. Accordingly, we devise the feedback objective for the encoder to follow the IoU map.

Similar to Eq. 3 in the decoder SA, we denote the IoU matrix $P^e_{QQ} \in \mathbb{R}^{N^e_q \times N^e_q}$ from the predictions of the encoder features, where N^e_q is the number of the encoder features. After normalizing P^e_{QQ} by a softmax function, the feedback objective with predictions for encoder SA is defined as

$$\mathcal{L}^{\text{e}}_{\text{SA}} = D_{KL}(A^e \,||\, P^e_{QQ}). \quad (6)$$

On the other hand, the CA map contains not only the query relation but also the relation of the encoder features. As such, we extend the prediction-feedback for the CA as described in Eq. 4. Specifically, the feature relation A^c_{KK} is

extracted from the CA by a matrix multiplication, similarly to A^c_{QQ}. Here, A^c_{KK} represents the similarities between encoder features to which similar groups of decoder queries attend. To conclude, $\mathcal{L}^{\text{d}}_{\text{CA}}$ defined in Eq. 4 is strengthened to

$$\mathcal{L}^{\text{d}}_{\text{CA}} = D_{KL}(A^c_{QQ} \,||\, P^d_{QQ}) + D_{KL}(A^c_{KK} \,||\, P^e_{QQ}). \quad (7)$$

Discussion. During the initial training phases, the model generates undertrained predictions. One might be concerned that the early feedback harms the learning of the model. However, in the initial iterations, the objective of TAD is primarily optimized over the feedback, ensuring that undertrained feedback does not disrupt the training. Additionally, note that the guidance derived from predictions does not constitute the optimal relation for attention. The feedback acts as a regularizer, helping the attention maps stay close to the predictions and maintain their balance with the main objective. Meanwhile, when the prediction-feedback relieves the collapse, the soundness of the CA is restored. This brings about the restoration of the full functionality of the previous work, Self-DETR. Experimental results demonstrate that the recovered CA remarkably boosts its performance.

Objectives

DETR. Let us denote the ground-truths, and the M predictions as y, $\hat{y} = \hat{y}_i{}_{i=1}^{M}$, respectively. For the bipartite matching between the ground-truth and prediction sets, the optimal matching is defined to search for the permutation of M elements $j \in J_M$ with the minimal cost as below:

$$\hat{j} = \underset{j \in J_M}{\arg\min} \sum_{i}^{M} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{j(i)}), \quad (8)$$

where $L_{\text{match}(y_i, \hat{y}_{j(i)})}$ is a pair-wise matching cost between y_i and the prediction with the index from $j(i)$, which produces the index i from the permutation j.

Next, we denote each ground-truth action as $y_i = (c_i, t_i)$, where c_i is the target class label with the background category $\emptyset$ and t_i is the time intervals of the start and end times. For the prediction with the index $j(i)$, we define the probability of the class c_i as $\hat{p}_{j(i)}(c_i)$ and the predictions of the time intervals as $\hat{t}_{\hat{j}(i)}$. Then $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{j(i)})$ is defined as below:

$$\mathcal{L}_{\text{match}}(y_i, \hat{y}_{j(i)}) = -\mathbb{1}_{c_i \neq \emptyset}\, \hat{p}_{j(i)}(c_i) + \mathbb{1}_{c_i \neq \emptyset}\, \mathcal{L}_{\text{reg}}(t_i, \hat{t}_{j(i)}),$$

where $\mathcal{L}_{\text{reg}}(t_i, \hat{t}_{j(i)})$ is the regression loss between the ground-truth t_i and the prediction $\hat{t}$ with the index $j(i)$. The regression loss $\mathcal{L}_{\text{reg}}$ consists of L1 and Interaction-over-Union (IoU) losses as in the DETR-based methods. Finally, we formulate the main objective as follows:

$$\mathcal{L}_{\text{DETR}}(y, \hat{y}) = \sum_{i=1}^{M} [-\log \hat{p}_{\hat{j}(i)}(c_i) + \mathbb{1}_{c_i \neq \emptyset} \mathcal{L}_{\text{reg}}(t_i, \hat{t}_{\hat{j}(i)})], \quad (9)$$

where $\hat{j}$ is the optimal assignment from Eq. 8.

Full Objectives. To summarize the objectives for our framework, Pred-DETR, the full objective is can be described as below:

$$\mathcal{L} = \mathcal{L}_{\text{DETR}} + \lambda^e_{\text{SA}} \mathcal{L}^{\text{e}}_{\text{SA}} + \lambda^d_{\text{SA}} \mathcal{L}^{\text{d}}_{\text{SA}} + \lambda^d_{\text{CA}} \mathcal{L}^{\text{d}}_{\text{CA}}, \quad (10)$$

where λ^e_{SA}, λ^d_{SA} and λ^d_{CA} are the weights for the prediction-feedback losses for the encoder and decoder.

Experiments

Datasets

In this paper, we utilize four challenging benchmarks of temporal action detection: THUMOS14 (Jiang et al. 2014), ActivityNet-v1.3 (Fabian Caba Heilbron and Niebles 2015), HACS (Zhao et al. 2019) and FineAction (Liu et al. 2022c). **THUMOS14** has 200 and 213 videos for the training and validation sets, respectively. The dataset has 20 action classes related to sports. **ActivityNet-v1.3** contains 19,994 videos with 200 action classes. 10024, 4926, and 5044 videos are for training, validation, and testing, respectively. **HACS** contains 37613 and 5981 videos are for training, validation, respectively, with 200 classes, shared by ActivityNet-v1.3. **FineAction** contains daily events with 106 categories and 16732 videos. THUMOS14 and FineAction contain many short actions while a majority of videos in ActivityNet-v1.3 and HACS have long actions.

Implementation Details

In this section, we briefly deliver the implementation details. For the detailed description, we recommend to refer to the supplementary materials.

Architecture. We utilize the features of I3D (Carreira and Zisserman 2017) pre-trained on Kinetics (Kay et al. 2017) for THUMOS14 and ActivityNet-v1.3. Also, we adopt SlowFast (Feichtenhofer et al. 2019) and VideoMAEv2-g (Wang et al. 2023) for HACS and FineAction, respectively. We are based on a temporal version of DAB-DETR as in Self-DETR. The number of layers of the encoder and decoder is 2, and 4, respectively. The number of the queries is 40. We set the weights λ^e_{SA}, λ^d_{SA} and , λ^d_{CA} of the losses of the prediction-feedback for the encoder and decoder as 2.

Enhanced DAB-DETR for TAD. We also introduce advanced tricks on DAB-DETR including stable matching (Liu et al. 2023), hybrid matching (Jia et al. 2023) and the two-stage mechanism from Deformable-DETR. Stable matching utilizes the IoU value between the prediction and the ground-truth as the target value of the class probability. It is closely related to the actionness regression in TadTR. Note that we do not utilize the predictions from the encoder as the initial decoder queries. We found that stable matching remarkably improves the performance, aligned with the results of TadTR. However, the two-stage mechanism slightly improves it because it is introduced for the prediction-feedback. We also report a study for the benefits from each component in the supplementary materials.

Main Results

Comparison with the State-of-the-Art. Table. 1 shows the comparison results with the state-of-the-art methods on THUMOS14 and ActivityNet-v1.3. Furthermore, Table. 2 and Table. 3 show the comparison results on HACS and FineAction. Pred-DETR outperforms DETR-based methods over various benchmarks. The first section denoted by 'Standard Methods' contains non-DETR approaches, and the second one includes DETR-based models. Also, in the DETR-based models, RTD-Net, Self-DETR, and ours are based on

Method	THUMOS14							ActivityNet-v1.3				
	Feat.	0.3	0.4	0.5	0.6	0.7	Avg.	Feat.	0.5	0.75	0.95	Avg.
Standard Methods												
BMN (Lin et al. 2019)	TSN	56.0	47.4	38.8	29.7	20.5	38.5	TSN	50.07	34.78	8.29	33.85
G-TAD (Xu et al. 2020)	TSN	54.5	47.6	40.2	30.8	23.4	39.3	TSN	50.36	34.60	9.02	34.09
AFSD (Lin et al. 2021)	I3D	67.3	62.4	55.5	43.7	31.1	52.0	I3D	52.40	35.30	6.50	34.40
TAGS (Nag et al. 2022)	I3D	68.6	63.8	57.0	46.3	31.8	52.8	I3D	56.30	36.80	**9.60**	36.50
ActionFormer (Zhang, Wu, and Li 2022)	I3D	82.1	77.8	71.0	59.4	43.9	66.8	I3D	53.50	36.20	8.20	35.60
TriDet (Shi et al. 2023)	I3D	83.6	**80.1**	**72.9**	**62.4**	47.4	**69.3**	TSP	54.70	38.00	8.40	36.80
DyFaDet (Yang et al. 2024)	I3D	**84.0**	**80.1**	72.7	61.1	**47.9**	69.2	TSP	**58.10**	**39.60**	8.40	**38.50***
DETR-based Methods												
RTD-Net (Tan et al. 2021)	I3D	68.3	62.3	51.9	38.8	23.7	49.0	I3D	47.21	30.68	8.61	30.83
TadTR (Liu et al. 2022b)	I3D	74.8	69.1	60.1	46.6	32.8	56.7	I3D	52.83	37.05	**10.83**	36.11
ReAct (Shi et al. 2022)	TSN	69.2	65.0	57.1	47.8	35.6	55.0	TSN	49.60	33.00	8.60	32.60
Self-DETR (Kim, Lee, and Heo 2023)	I3D	74.6	69.5	60.0	47.6	31.8	56.7	I3D	52.25	33.67	8.40	33.76
Pred-DETR (Ours)	I3D	80.0	73.5	64.6	52.3	37.6	61.6	I3D	54.17	36.43	9.53	36.00
Pred-DETR (Ours)	VM2	**84.1**	**80.0**	**72.2**	**60.4**	**45.8**	**68.5**	I3D	**58.38**	**39.14**	9.92	**38.62***

Table 1: **The comparison results with the state-of-the-art on THUMOS14 and ActivityNet-v1.3.** 'Feat.' indicates the backbone features including I3D, TSN, TSP (R(2+1)D) and VM2. '*' means using UniFormer-v2 classification on ActivityNet-v1.3.

Method	Feat.	0.5	0.75	0.95	Avg.
Standard Methods					
G-TAD (Xu et al. 2020)	I3D	41.1	27.6	88.3	27.5
BMN (Lin et al. 2019)	SF	52.5	36.4	10.4	35.8
TCANet (Qing et al. 2021)	SF	54.1	37.2	11.3	36.8
TriDet (Shi et al. 2023)	SF	**56.7**	**39.3**	**11.7**	**38.6**
DETR-based Methods					
TadTR (Liu et al. 2022b)	I3D	47.1	32.1	10.9	32.1
Self-DETR† (Kim et al. 2023)	I3D	49.9	31.1	9.3	31.8
Pred-DETR (Ours)	I3D	51.5	32.7	10.2	33.3
Pred-DETR (Ours)	SF	**56.5**	**36.8**	**12.1**	**37.4**

Table 2: **The comparison results on the HACS dataset.** '†' indicates our reproduced version.

Method	Feat.	0.5	0.75	0.95	Avg.
Standard Methods					
DBG (Lin et al. 2020)	I3D	10.7	6.4	2.5	6.8
BMN (Lin et al. 2019)	I3D	13.7	8.8	3.1	9.1
G-TAD (Xu et al. 2020)	I3D	14.4	8.9	3.1	9.3
ActionFormer (Zhang et al. 2022)	VM2	**29.1**	**17.7**	**5.1**	**18.2**
DETR-based Methods					
TadTR† (Liu et al. 2022b)	VM2	21.3	8.7	0.4	10.3
Self-DETR† (Kim et al. 2023)	VM2	22.1	9.2	0.4	10.7
Pred-DETR (Ours)	VM2	**25.9**	**12.2**	**1.0**	**13.4**

Table 3: **The comparison results on the FineAction dataset.** '†' indicates our reproduced version.

standard attention. Also, TadTR and ReAct are based on deformable attention. We also indicate the backbone features by 'Feats'. Most approaches utilize the TSN (Wang et al. 2016) or I3D features while some methods also adopt the TSP (Alwassel, Giancola, and Ghanem 2021) features.

In the table, our model remarkably outperforms all DETR-based models on all benchmarks. It demonstrates that when the attention collapse is relieved, the original-DETR architecture can be comparable or superior to the Deformable-DETR architecture in TAD, aligned with the observation in object detection (Lin et al. 2023). More interestingly, Pred-DETR best performs on ActivityNet-v1.3 including the non-DETR methods. DETR-based methods tend to produce better results on ActivityNet and HACS than those on THUMOS14 and FineAction. This could be because ActivityNet and HACS mostly contain long actions while THUMOS14 and FineAction include many short instances. Predicting short actions precisely requires high temporal resolution while DETR does not handle yet such a long sequence due to the query-based architecture. Nevertheless, recent DETR models including ours show superior performances over the non-DETR models which handle short-length sequences except for ActionFormer and TriDet.

Analysis

Diversity. To clearly validate the effect of the feedback for the collapse, we measure the diversity of cross- and self-attention maps according to (Dong, Cordonnier, and Loukas 2021; Kim, Lee, and Heo 2023). The diversity $d(A)$ for the attention map A is the measure of the closeness between the attention map and a rank-1 matrix as below:

$$d(A) = \|A - \mathbf{1}a^\top\| \,/\, \|A\|, \text{where } a = \arg\min_{a'} \|A - \mathbf{1}a'^\top\|,$$

where $\|\cdot\|$ denotes the ℓ_1,ℓ_∞-composite matrix norm, a, a' are column vectors of the attention map A, and $\mathbf{1}$ is an all-ones vector. Note that the rank of $\mathbf{1}a^\top$ is 1, and therefore, a smaller value of $d(A)$ means A is closer to a rank-1 matrix.

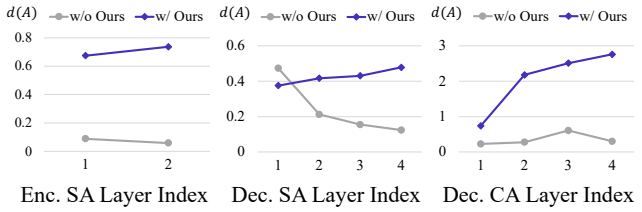

Figure 4: **Diversity of attention maps.** Diversity for cross- and self-attention for test samples of ActivityNet-v1.3.

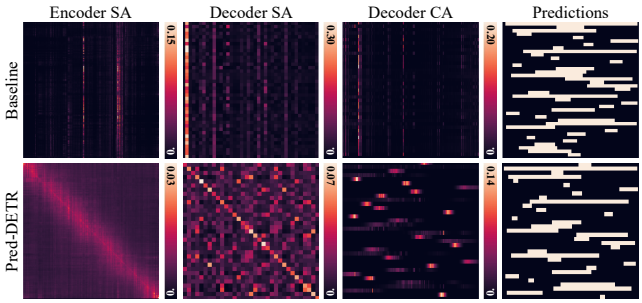

Figure 5: **Attention maps.** The figure shows self- and cross-attention maps from a test sample in ActivityNet-v1.3.

Fig. 4 shows the diversity on each layer of the encoder and decoder for the baseline DETR and Pred-DETR. The diversity is measured on the test set on ActivityNet-v1.3 averaged over all test samples. As the model depth gets deeper, the diversity of the baseline decreases close to 0. However, the diversity of Pred-DETR does not fall down and even increases. From this, Pred-DETR effectively relieves the collapse.

Attention Maps. Fig. 5 shows the visualization of the self- and cross-attention maps from the encoder and decoder. As shown, the baseline DETR exhibits the attention collapse in all attention modules. On the other hand, our model does not suffer from the collapse and shows expressive attention.

Ablation on Prediction-Feedback. In order to validate the benefits from each component in our framework, we conducted the ablation study on the self-feedback objectives. In Pred-DETR, we have three types of feedback for 1) decoder cross-attention ($\mathcal{L}^{\text{d}}_{\text{CA}}$), 2) decoder self-attention ($\mathcal{L}^{\text{d}}_{\text{SA}}$), and 3) encoder self-attention ($\mathcal{L}^{\text{e}}_{\text{SA}}$).

Table. 4 shows the ablation results. As shown, each type of prediction-feedback clearly improves the performance. Also, when we introduce all three kinds of prediction-feedback, the benefits become the most. The prediction-feedback for the cross-attention modules brings the most performance gain as they are the central part of DETR.

Prediction-Feedback Targets. As for the targets for the self-attention in feedback, we can also adopt the guidance map from the cross-attention proposed in Self-DETR (Kim, Lee, and Heo 2023). The upper of Table. 5 shows the results with Self-DETR. When we do not use the prediction-feedback for the cross-attention, we can see that the the feedback from the predictions (denoted by 'Pred Relation' in the table) show superior performance to that from the cross-attention (From CA). Also, when Self-DETR is introduced with our prediction-feedback for the cross-attention, the performance gain becomes a way larger because the attention collapse of the cross-attention is remarkably relieved.

$\mathcal{L}^{\text{e}}_{\text{SA}}$	$\mathcal{L}^{\text{d}}_{\text{SA}}$	$\mathcal{L}^{\text{d}}_{\text{CA}}$	THUMOS14				ActivityNet-v1.3			
			0.3	0.5	0.7	Avg.	0.5	0.75	0.95	Avg.
·	·	·	73.8	57.0	26.5	53.5	52.65	33.79	8.93	34.14
✓	·	·	79.7	64.0	32.0	59.7	56.91	36.27	9.59	36.67
·	✓	·	77.0	61.9	34.6	58.8	54.60	34.65	9.25	35.11
·	·	✓	79.0	64.1	36.1	60.8	57.59	38.64	9.81	38.14
✓	✓	·	79.0	63.3	35.4	60.5	57.63	37.88	9.46	37.65
✓	✓	✓	80.0	64.6	37.6	61.6	58.38	39.14	9.92	38.62

Table 4: **Ablation on prediction-feedback.** The ablation study is conducted on THUMOS14 and ActivityNet-v1.3.

Cross-Attn.	Self-Attn.	0.5	0.75	0.95	Avg.
-	From CA	56.14	36.10	9.12	36.18
-	Pred Relation	57.63	37.88	9.46	37.65
Pred Relation	From CA	57.80	38.57	10.16	38.32
Pred Relation	Pred Relation	58.38	39.14	9.92	38.62
Ground-Truth	Pred Relation	53.41	33.76	8.96	34.12
Pred Intervals	Pred Relation	53.41	34.33	8.93	34.62
Pred Relation	Pred Relation	58.38	39.14	9.92	38.62

Table 5: **Prediction-feedback targets.** We conduct the study on the prediction-feedback targets for the cross- and self-attention on ActivityNet-v1.3.

In our prediction-feedback for cross-attention, we propose to utilize the indirect relation from the cross-attention. One may think a direct solution as the objective where the ground-truth or prediction intervals are matched to the cross-attention map. However, we claim that this way significantly harms the diversity of the representations for the CA mainly because we do not exactly know where the cross-attention should focus on. The bottom of Table. 5 shows the results with three types of the targets on ActivityNet-v1.3. The targets from the intervals of the ground-truth or predictions ('Ground-Truth' and 'Prediction Intervals', respectively) degrade the performance as expected. However, the indirect way with the relation of the predictions (Prediction Relation) remarkably improves the performance.

Conclusion

In this paper, we discovered the attention collapse in the cross-attention of DETR for TAD. We found that the model exhibits a clearly different pattern from the predictions, which is a short-cut phenomenon from the collapse. To this end, we proposed Prediction-Feedback DETR (Pred-DETR) to align the attention with the predictions. By providing an auxiliary objective with the guidance from the predictions, the prediction-feedback remarkably relieved the degree of the collapse. Our extensive experiments demonstrated that Pred-DETR recovered the diversity of the attention with the state-of-the-art performance over DETR models on THUMOS14, ActivityNet-v1.3, HACS, and FineAction.

Acknowledgements

This work was supported in part by MSIT/IITP (No. 2022-0-00680, 2020-0-01821, 2019-0-00421, RS-2024-00459618, RS-2024- 00360227, RS-2024-00437102, RS-2024-00437633), and MSIT/NRF (No. RS-2024-00357729).

References

Alwassel, H.; Giancola, S.; and Ghanem, B. 2021. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3173–3183.

Buch, S.; Escorcia, V.; Shen, C.; Ghanem, B.; and Carlos Niebles, J. 2017. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2911–2920.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Dai, X.; Singh, B.; Zhang, G.; Davis, L. S.; and Qiu Chen, Y. 2017. Temporal context network for activity localization in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 5793–5802.

Dong, Y.; Cordonnier, J.-B.; and Loukas, A. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, 2793–2803. PMLR.

Fabian Caba Heilbron, B. G., Victor Escorcia; and Niebles, J. C. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–970.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gao, J.; Yang, Z.; Chen, K.; Sun, C.; and Nevatia, R. 2017. Turn tap: Temporal unit regression network for temporal action proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, 3628–3636.

Jia, D.; Yuan, Y.; He, H.; Wu, X.; Yu, H.; Lin, W.; Sun, L.; Zhang, C.; and Hu, H. 2023. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19702–19712.

Jiang, Y.-G.; Liu, J.; Roshan Zamir, A.; Toderici, G.; Laptev, I.; Shah, M.; and Sukthankar, R. 2014. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kim, J.; Choi, J.; Jeon, Y.; and Heo, J.-P. 2024. Boundary-Recovering Network for Temporal Action Detection. *arXiv preprint arXiv:2408.09354*.

Kim, J.; and Heo, J. 2019. Learning Coarse and Fine Features for Precise Temporal Action Localization. *IEEE Access*, 7: 149797–149809.

Kim, J.; Lee, M.; and Heo, J.-P. 2023. Self-Feedback DETR for Temporal Action Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10286–10296.

Kim, J.; Lee, M.; and Heo, J.-P. 2024. Long-Term Pretraining for Temporal Action Detection with Transformers. *arXiv preprint arXiv:2408.13152*.

Lin, C.; Li, J.; Wang, Y.; Tai, Y.; Luo, D.; Cui, Z.; Wang, C.; Li, J.; Huang, F.; and Ji, R. 2020. Fast learning of temporal action proposal via dense boundary generator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11499–11506.

Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Fu, Y. 2021. Learning Salient Boundary Feature for Anchor-free Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3320–3329.

Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3889–3898.

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

Lin, Y.; Yuan, Y.; Zhang, Z.; Li, C.; Zheng, N.; and Hu, H. 2023. Detr does not need multi-scale or locality design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6545–6554.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022a. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Liu, S.; Ren, T.; Chen, J.; Zeng, Z.; Zhang, H.; Li, F.; Li, H.; Huang, J.; Su, H.; Zhu, J.; et al. 2023. Detection transformer with stable matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6491–6500.

Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; and Bai, X. 2022b. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441.

Liu, Y.; Wang, L.; Wang, Y.; Ma, X.; and Qiao, Y. 2022c. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31: 6937–6950.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training

convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3651–3660.

Nag, S.; Zhu, X.; Song, Y.-Z.; and Xiang, T. 2022. Proposal-free temporal action detection via global segmentation mask learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, 645–662. Springer.

Qing, Z.; Su, H.; Gan, W.; Wang, D.; Wu, W.; Wang, X.; Qiao, Y.; Yan, J.; Gao, C.; and Sang, N. 2021. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 485–494.

Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; and Tao, D. 2023. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18857–18866.

Shi, D.; Zhong, Y.; Cao, Q.; Zhang, J.; Ma, L.; Li, J.; and Tao, D. 2022. 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*, 105–121. Springer.

Shou, Z.; Wang, D.; and Chang, S.-F. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1049–1058.

Su, H.; Gan, W.; Wu, W.; Qiao, Y.; and Yan, J. 2021. Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2602–2610.

Tan, J.; Tang, J.; Wang, L.; and Wu, G. 2021. Relaxed Transformer Decoders for Direct Action Proposal Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13526–13535.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14549–14560.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, 20–36. Springer.

Xu, H.; Das, A.; and Saenko, K. 2017. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, 5783–5792.

Xu, M.; Zhao, C.; Rojas, D. S.; Thabet, A.; and Ghanem, B. 2020. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10156–10165.

Yang, L.; Zheng, Z.; Han, Y.; Cheng, H.; Song, S.; Huang, G.; and Li, F. 2024. DyFADet: Dynamic Feature Aggregation for Temporal Action Detection. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, volume 15104 of *Lecture Notes in Computer Science*, 305–322. Springer.

Yeung, S.; Russakovsky, O.; Mori, G.; and Fei-Fei, L. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2678–2687.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 492–510. Springer.

Zhao, H.; Yan, Z.; Torresani, L.; and Torralba, A. 2019. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. *arXiv preprint arXiv:1712.09374*.

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2914–2923.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.