A. Supplementary Material

A.1. Network Architecture in Feature Pyramid

From Transformer to CNN. To be self-contained, we analyze the impact of module design on the detector. For comparison, we build two baseline models: a convolutional baseline and a Transformer baseline. Firstly, we build the convolution baseline where the convolutional module is adopted from the previous one-stage detector [21, 49]. Secondly, the previous state-of-the-art detector [49] with the local window self-attention [4] is chosen as the Transformer baseline. Then, to analyze the importance of two common components: self-attention and normalization, in the Transformer [41] macrostructure, we provide three variants of the convolutional-based structure: SA-to-CNN, LN-to-GN and LN-GN-Mix, as Fig. 7 shown, and validate their performance on THUMOS14.

Results. From the Tab. 9, we can see there is a large performance gap between the Transformer baseline and the CNN baseline (about 8.1% in average mAP), demonstrating that the Transformer holds a large advantage for TAD tasks. Then, we conduct the ablation study with the three variants with normal regression head and Trident-head, respectively.

We first simply replace the local self-attention with a 1D convolutional layer which has the same receptive field with [49] (*e.g.* kernel size is 19). This change brings a dramatic performance increase in average mAP compared with the CNN baseline (about 6.2%) but is still behind the Transformer baseline by about 1.9%. Next, we conduct experiments with different normalization layers (*i.e.* Layer Normalization (LN) [3] and Group Normalization (GN) [43]) and we find that the hybrid structure of LN and GN (LN-GN-Mix) shows better performance comparing to the original form of the Transformer (65.7 versus 64.9). By combining with the Trident-head, the LN-GN-Mix version achieves 66.0% in average mAP, which demonstrates the possibility of efficient convolutional modeling. These empirical results further motivate us to improve the feature pyramid with SGP layer (see Sec 3.2 of the main test for more details).

A.2. The rank loss problem in Transformer.

In [13], the authors discuss how the pure self-attention operation causes the input feature to converge to a rank-1 matrix at a double exponential rate, while MLP and residual connections can only partially slow down this convergence. This phenomenon is disastrous for TAD tasks, as the video feature sequences extracted by pre-trained action recognition networks are often highly similar (see Section 1), which further aggravates the rank loss problem and makes the features at each instant indistinguishable, resulting in inaccurate detection of action.

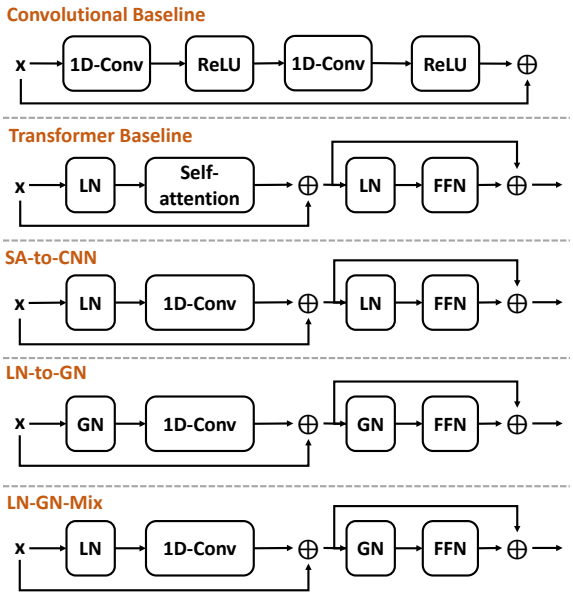

Figure 7. Two baseline models and three different variants of the convolutional-based structure.

Table 9. The results of different variants on THUMOS14. *: with Trident-head.

Method	0.3	0.4	0.5	0.6	0.7	Avg.
CNN Baseline	73.7	68.8	61.4	51.6	38.0	58.7
Transformer Baseline	82.1	77.8	71.0	59.4	43.9	66.8
SA-to-CNN	80.4	76.4	67.5	57.5	42.9	64.9
LN-to-GN	80.0	76.3	68.0	57.2	42.3	64.8
LN-GN-Mix	80.8	77.2	68.8	58.1	43.6	65.7
SA-to-CNN*	81.2	77.3	68.7	58.0	43.5	65.7
LN-to-GN*	80.7	76.9	69.1	58.0	42.2	65.4
LN-GN-Mix*	81.6	77.7	69.5	58.2	42.9	66.0

We posit that the core reason for this issue lies in the softmax function used in self-attention. Namely, the probability matrix (*i.e.* softmax(QK^T)) is *non-negative* and *the sum of each row is 1*, indicating the outputs of SA are *convex combination* for the value feature V. We will demonstrate that the largest angle between any two features in $V' = SA(V)$ is always less than or equal to the largest angle between features in V.

Definition A.2.1 (Convex Combination) *Given a set of points $S = \{x_1, x_2..., x_n\}$, a convex combination is a point of the form $\sum_n a_n x_n$, where $a_n \geq 0$ and $\sum_n a_n = 1$.*

Definition A.2.2 (Convex Hull) *The convex hull H of a given set of points S is identical to the set of all their convex combinations. A Convex hull is a convex set.*

Property A.2.2.1 (Extreme point) *An extreme point p is a point in the set that does not lie on any open line segment between any other two points of the same set. For a point set S and its convex hull H, we have $p \in S$.*

Lemma A.2.3 *Consider the case of a convex hull that does not contain the origin. Let $a, b \in \mathbb{R}^n$ and let S be the convex hull formed by them. Then, the angle between any two position vectors of points in S is less than or equal to the angle between the position vectors of the extreme points $\vec{a}$ and $\vec{b}$.*

Proof A.2.3.1 *Consider the objective function*

$$f(x) = \cos(\vec{x}, \vec{y}) = \frac{\langle \vec{x}, \vec{y} \rangle}{\|\vec{x}\|_2 \|\vec{y}\|_2},$$

where $\vec{x}, \vec{y}$ are the position vectors of two points x_1, x_2 within the convex hull S (a line segment with extreme points a and b). The angle between two vectors is invariant with respect to the magnitude of the vectors, thus, for simplicity, we define $\vec{x} = \vec{a} + x\vec{b}$, $\vec{y} = \vec{a} + y\vec{b}$, where $x, y \in [0, +\infty)$. Moreover, we have

$$\begin{aligned} f'(x) = &\|\vec{x}\|_2^{-3} \|\vec{y}\|_2^{-1} \times \\ &[\langle \vec{b}, \vec{y} \rangle ||\vec{a} + x\vec{b}||_2^2 - (||\vec{b}||_2^2 x + \langle \vec{a}, \vec{b} \rangle)\langle \vec{a} + x\vec{b}, \vec{y} \rangle] \end{aligned}$$

We consider

$$\begin{aligned} g(x) =&\langle \vec{b}, \vec{y} \rangle ||\vec{a} + x\vec{b}||_2^2 - (||\vec{b}||_2^2 x + \langle \vec{a}, \vec{b} \rangle)\langle \vec{a} + x\vec{b}, \vec{y} \rangle \\ =&\langle \vec{b}, \vec{y} \rangle (||\vec{a}||_2^2 + 2\langle \vec{a}, \vec{b} \rangle x + ||\vec{b}||_2^2 x^2) - [\langle \vec{b}, \vec{y} \rangle ||b||_2^2 x^2 \\ &+ (\langle \vec{a}, \vec{b} \rangle ||b||_2^2 + \langle \vec{a}, \vec{b} \rangle \langle \vec{b}, \vec{y} \rangle) x + \langle \vec{a}, \vec{y} \rangle \langle \vec{a}, \vec{b} \rangle] \\ =&(\langle \vec{a}, \vec{b} \rangle \langle \vec{b}, \vec{y} \rangle - \langle \vec{a}, \vec{y} \rangle \langle \vec{b}, \vec{b} \rangle) x + \langle \vec{a}, \vec{a} \rangle \langle \vec{b}, \vec{y} \rangle - \langle \vec{a}, \vec{y} \rangle \langle \vec{a}, \vec{b} \rangle. \end{aligned}$$

Substituting $\vec{y} = \vec{a} + y\vec{b}$ into the above equation, we have

$$\begin{aligned} g(x) =&(\langle \vec{a}, \vec{b} \rangle \langle \vec{b}, \vec{a} + y\vec{b} \rangle - \langle \vec{a}, \vec{a} + y\vec{b} \rangle \langle \vec{b}, \vec{b} \rangle) x + \\ &\langle \vec{a}, \vec{a} \rangle \langle \vec{b}, \vec{a} + y\vec{b} \rangle - \langle \vec{a}, \vec{a} + y\vec{b} \rangle \langle \vec{a}, \vec{b} \rangle \\ =&[\langle \vec{a}, \vec{b} \rangle (\langle \vec{a}, \vec{b} \rangle + y\langle \vec{b}, \vec{b} \rangle) - (\langle \vec{a}, \vec{a} \rangle + y\langle \vec{a}, \vec{b} \rangle)\langle \vec{b}, \vec{b} \rangle] x + \\ &[\langle \vec{a}, \vec{a} \rangle (\langle \vec{a}, \vec{b} \rangle + y\langle \vec{b}, \vec{b} \rangle) - (\langle \vec{a}, \vec{a} \rangle + y\langle \vec{a}, \vec{b} \rangle)\langle \vec{a}, \vec{b} \rangle] \\ =&(||\langle \vec{a}, \vec{b} \rangle||_2^2 - ||\vec{a}||_2^2 ||\vec{b}||_2^2) x + (||\vec{a}||_2^2 ||\vec{b}||_2^2 - ||\langle \vec{a}, \vec{b} \rangle||_2^2) y \\ =&(||\langle \vec{a}, \vec{b} \rangle||_2^2 - ||\vec{a}||_2^2 ||\vec{b}||_2^2)(x - y). \end{aligned}$$

According to the Cauchy-Schwarz inequality, we can obtain

$$||\langle \vec{a}, \vec{b} \rangle||_2^2 - ||\vec{a}||_2^2 ||\vec{b}||_2^2 \leq 0$$

Then, we have

$$g(x) \begin{cases} > 0 & x < y \\ = 0 & x = y \\ < 0 & x > y. \end{cases}$$

thus, for any position vector $\vec{y}$, when $x = 0$ or $x \to \infty$ ($\vec{x} = \vec{a}$ or $\vec{x} = \vec{b}$), the angle formed between $\vec{y}$ and $\vec{x}$ is maximum.

Without loss of generality, given a specific $\vec{y}$, if its maximum vector $\vec{x} = \vec{a}$, we can then set $\vec{y}$ to $\vec{a}$ and find its maximum vector again, which yields

$$\theta(\vec{x}, \vec{y}) \leq \theta(\vec{a}, \vec{y}) \leq \theta(\vec{b}, \vec{a})$$

The proof is completed.

Theorem A.2.4 *Consider the case of a convex hull that does not contain the origin. Let $X = \{x_1, x_2, \ldots, x_k\}$ be a set of points and let S be its convex hull. Then, the maximum angle between the position vectors of any two points in S is formed by the position vectors of two extreme points of S.*

Proof A.2.4.1 *Assume that this case holds when k.*

When $k = 2$, based on Lemma A.2.3, the maximum angle is formed by the extreme points $\vec{x_1}$ and $\vec{x_2}$.

When $k \geq 3$, we can sort the elements of X such that for a point y in S, $\vec{x_k}$ maximizes the angle $\theta(\vec{y}, \vec{x_k})$. Besides, the points x in S are of the form:

$$\begin{aligned} &\lambda_1 \vec{x_1} + \lambda_2 \vec{x_2} + ... + \lambda_k \vec{x_k} \\ =&(\lambda_1 + ... + \lambda_{k-1})(\frac{\lambda_1 \vec{x_1}}{\lambda_1 + ... + \lambda_{k-1}} + ... + \frac{\lambda_{k-1} \vec{x_{k-1}}}{\lambda_1 + ... + \lambda_{k-1}}) \\ &+ \lambda_k \vec{x_k}, \end{aligned}$$

where $(\frac{\lambda_1 \vec{x_1}}{\lambda_1 + ... + \lambda_{n-1}} + ... + \frac{\lambda_{k-1} \vec{x_{k-1}}}{\lambda_1 + ... + \lambda_{k-1}})$ is a position vector of a point located within the convex hull induced by $\{x_1, x_2, ..., x_{k-1}\}$. Through Lemma A.2.3 and definition, we can obtain

$$\theta(\vec{x}, \vec{y}) \leq \theta(\vec{x_k}, \vec{y})$$

For any two points x and y in a convex hull S, by setting $\vec{y} = \vec{x_k}$ and using the above inequality twice, without loss of generality, we can assume that the vector $\vec{x_1}$ makes the largest angle with $\vec{x_k}$. Then, we can obtain

$$\theta(\vec{x}, \vec{y}) \leq \theta(\vec{x_k}, \vec{y}) \leq \theta(\vec{x_1}, \vec{x_k})$$

By definition, $\theta(\vec{x_1}, \vec{x_k})$ is no greater than the maximum angle formed by any other two basis vectors.

The proof is completed.

Corollary A.2.5 *When the convex hull of the input set V does not contain the origin, the largest angle between any two features after self-attention $V' = SA(V)$ is always less than or equal to the largest angle between features in V.*

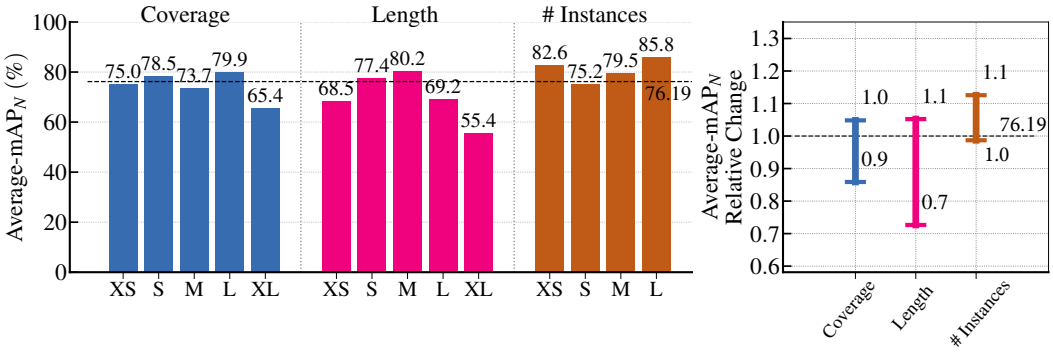

Figure 8. The sensitivity analysis of the detection results on THUMOS14. Where mAP_N is the normalized mAP with the average number N of ground truth segments per class [2].

Remark A.2.5.1 *In the Temporal Action Detection (TAD) task, the temporal feature sequences extracted by the pre-trained video classification backbone often exhibit high similarity and the pure Layer Normalization [3] projects the input features onto the hyper-sphere in the high-dimensional space. Consequently, the convex hull induced by these features often does not encompass the origin. As a result, self-attention operation causes the input features to become more similar, reducing the distinction between temporal features and hindering the performance of the TAD task.*

A.3. Error Analysis

In this section, we analyse the detection results on THUMOS14 with the tool from [2], which analyze the results in three main directions: the False Positive (FP), the False Negative (FN) and the sensitivity of different length. For a further explanation of the analysis, please refer to [2] for more details.

Sensitivity analysis. As shown in Fig. 8 (Left), three metrics are taken into consideration: coverage (the normalized length of the instance by the duration of the video), length (the actual length in seconds) and the number of instances (in a video). The results are divided into several length/number bins from extremely short (XS) to extremely long (XL). We can see that our method's performance is balanced over most of the action length, except for extremely long action instances which are significantly lower than the overall value (the dashed line). That's because extremely long action instances contain more complicated information, which deserves further exploration.

Analysis of the false positives. Fig. 9 shows a chart of the percentage of different types of action instances in different $k - G$ numbers, where G is the number of the ground-truth instances for each action category and the top $k \times G$ predicted instances are kept for visualization.

From the $1G$ column on left, we can see in the top G prediction, the true positive instances account for about 80% (at

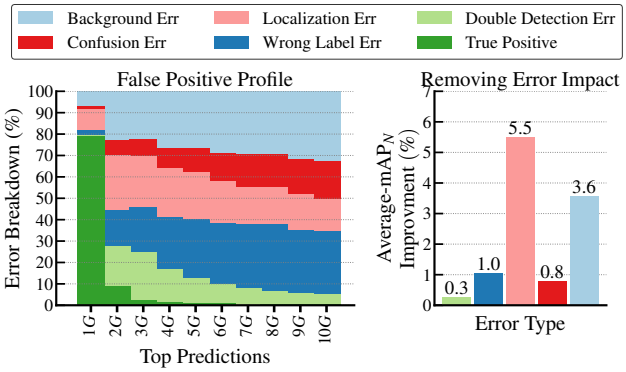

Figure 9. The false positive profile. It counts the percentage of several common types of detection error in different Top-K prediction groups.

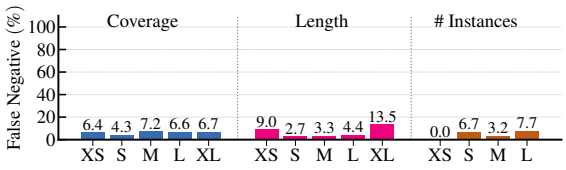

Figure 10. The false negative profile. It counts the percentage of miss-detection instances in different video lengths or videos with different action instance densities.

IoU=0.5), which indicates that our method has the power to estimate the right score of each instance. Moreover, on right, we can see the impact of each type of error: the regression error (*i.e.* localization error and background error, the IoU between prediction and ground truth is much lower than a threshold or equal to zero) is still the part that deserves the most attention.

Analysis of the false negatives. In this section, we analyze the false negative (miss-detection) rate for our method. As depicted in Fig. 10, only the extremely short and extremely

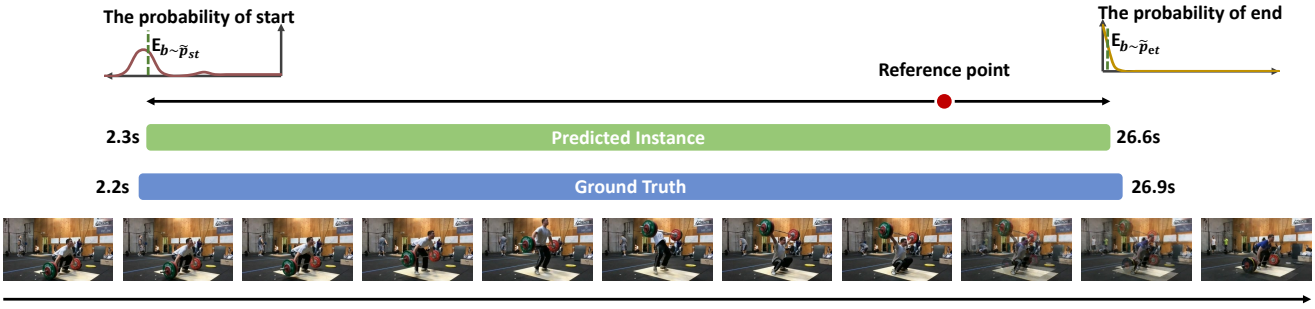

Figure 11. A visualization of the detection result on the THUMOS14 test set.

long instances have a relatively higher FN rate (9.0% and 13.5%, respectively), which is consistent with intuition that they are more difficult to detect. Note that for a video with only one action instance (XS), TriDet can detect all of them without any miss-detection (0.0 in # Intances), demonstrating our advantage for single-action localization.

A.4. Qualitative Analysis

In Fig. 11, we show the visualization of a detection result on the THUMOS14 test set. It can be seen that our method accurately predicts the start and end instant of the action. Besides, we also visualize the predicted probability of the boundary in the Trident-head, where only the bin around the boundary has a relatively high probability while the others are low and smooth, indicating that the Trident-head can converge to a reasonable result.