Supplementary Material of Bridge the Gap: From Weak to Full Supervision for Temporal Action Localization with PseudoFormer

Ziyi Liu
University of Science and Technology Beijing
liuziyi.iair@gmail.com

Yangcen Liu
Georgia Institute of Technology
yliu3735@gatech.edu

i. Details on Full Branch Design

The overall regression-based model has a similar structure to TriDet [2] and AFormer [3]. With input snippet-level feature $\boldsymbol{F} \in \mathbb{R}^{N\times D}$, a transformer encoder would encode them as $\boldsymbol{\xi} \in \mathbb{R}^{T\times d}$, where d is the dimension.

Attention Head. In order to learn from the snippet-level predictions (SPs) $\boldsymbol{Z}$ as prior with the base model, we apply attention head with one convolutional layer followed by three stacked MLP layers to enhance the encoded representation $\boldsymbol{\xi}$. After getting the predicted logit $\boldsymbol{l}$, we apply a classification as:

$$\bar{\boldsymbol{Z}} = Softmax(\boldsymbol{l}), \tag{1}$$

where $\bar{\boldsymbol{Z}} \in \mathbb{R}^{T\times C+1}$ is the predicted SPs. We use L_{att} to train attention head.

FPN and Label Assignment. Utilizing a Feature Pyramid Network (FPN), the encoded features $\boldsymbol{\xi}$ are progressively down-sampled across multiple layers, serving as anchors at each level. Subsequently, all $\hat{T}$ anchors from different layers are concatenated. Each anchor predicts an action score for all classes through the classification head, along with corresponding temporal boundaries via the regression head, which is then used to decode action candidates. For label assignment, each proposal in $\hat{P}$ is assigned a duration and allocated to the appropriate FPN level based on its regression range. The architecture, combined with the label assignment strategy, ensures consistent handling of proposals across varying durations.

Regression Head. For the regression head, a 3-layer MLP is utilized to decode each anchor to a start offset $\bar{d}^f_{st}$ and end offset $\bar{d}^f_{ed}$, in the f layer. The predicted boundaries are calculated with:

$$\bar{s}_t = (t - \bar{d}^f_{st}) \times 2^{f-1}, \tag{2}$$

$$\bar{e}_t = (t + \bar{d}^f_{ed}) \times 2^{f-1}, \tag{3}$$

where $\bar{s}_t$ and $\bar{e}_t$ are separately the start and the end. L_{reg} is applied to train the regression head.

Methods	mAP@IoU(%)							AVG
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	(0.1:0.7)
Hard	72.1	66.7	58.2	48.0	36.4	25.9	14.7	46.0
Soft	75.9	70.2	61.8	50.9	38.7	25.6	14.6	48.2
Top-K	75.5	69.7	60.9	50.0	38.1	26.4	15.2	47.3
Threshold	75.0	69.9	61.4	51.6	40.6	28.1	16.3	49.0
Gauss	75.4	70.1	61.6	51.0	40.1	27.6	15.8	49.2
RickerFusion	**76.3**	**71.8**	**63.6**	**53.8**	**42.5**	**29.7**	**16.5**	**50.8**

Table 1. Comparison of different pseudo label generation strategies, as the reference of Tab. 2

Methods	mAP@IoU(%)							AVG
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	(0.1:0.7)
Hard	11.1	9.2	7.3	6.0	5.3	3.8	2.5	6.5
Soft	71.5	66.2	56.5	47.7	40.5	27.2	15.3	46.4
Top-K	56.1	51.4	43.5	36.5	30.2	19.6	10.6	35.4
Threshold	70.6	65.4	55.8	46.9	39.6	26.2	14.4	45.6
Gauss	69.5	64.2	54.6	45.7	38.3	25.8	13.8	43.7
RickerFusion	61.6	56.9	48.4	40.7	33.8	22.2	11.9	39.4

Table 2. Comparison of different pseudo label generation strategies **applied as postprocessing**. Compared to Tab. 1, the results show no clear order or consistent pattern in performance.

Classification Head. For the classification head, we retain the structure as a 3-layer MLP, consistent with the regression head. For each anchor, the regression head performs individual (C) 2-classification predictions to distinguish between the foreground and background. The classification head is optimized using L_{cls}.

ii. What Are Good Pseudo Labels?

In the context of two-branch methods, such as those explored in [1] and [4], an important question arises: what constitutes high quality pseudo labels?

Based on the observations in Tab. 2, where different pseudo-label generation strategies are directly used for post-processing before evaluating the results, we discovered that the performance of these methods is completely decoupled from the performance achieved by training a full branch with pseudo labels. This discrepancy is largely due to the evaluation mechanism, where metrics of mean Average Precision (mAP) often consider a large number of redundant proposals. Consequently, post-processing methods, such as the *Hard* or *Top-K* strategy, only retain a small subset of predictions (after *Hard* strategy, only 4% of predicted proposals are preserved), which can significantly degrade the overall performance.

This phenomenon stems from the fact that pseudo proposals require a clear boundary for each proposal to train the regression-based model. Specifically, each snippet should be assigned a binary label as a foreground or background. Furthermore, these pseudo labels should be calculated by considering the boundaries and scores of all output proposals, providing a comprehensive evaluation. This is why methods like *Gauss* and *RickerFusion* produce superior pseudo labels. They are better equipped to ensure clear boundaries and provide a more reliable fusion of proposal information.

$\beta\backslash\alpha$	0.00	0.05	0.10	0.15
0.00	51.5	51.8	51.9	51.4
0.05	51.1	51.3	51.2	51.0
0.10	50.4	50.8	50.7	50.5

Table 3. The average mAP (0.1:0.7) values for different values of α and β on the THUMOS14 dataset.

iii. Training with Noisy Labels.

Without proposal-level annotation, it is unavoidable that the pseudo labels could be noisy. In PseudoFormer, we apply an uncertainty mask and refinement strategy to deal with noisy label training. Also, in L_{att}, we apply threshold τ to filter the uncertain attention labels. The main challenge in training with noisy labels is effectively filtering out samples (e.g., snippets, anchors, proposals) with high uncertainty while retaining as many reliable samples as possible.

For the uncertainty mask, we vary the value of expansion ratio α and shrinking ratio β, and report the results on the THUMOS'14 dataset in Tab. 3. We report the performance by varying the value of α from 0.00 to 0.15, and β from 0.00 to 0.10. We do not use a larger value since we want to keep the number of snippets participating in training. With larger values α and β, more snippets around the boundaries with higher uncertainty are excluded while the total number for training decreases. For THUMOS14, we observe that performance declines as β increases, whereas the optimal results are achieved when α is set to 0.10. This indicates that the pseudo proposals for THUMOS14 exhibit higher confidence regarding the inner boundaries of actions but remain uncertain about their outer boundaries. For ActivityNet1.3, the two values are both 0.05 for the best performance, indicating that both sides of the boundaries are uncertain.

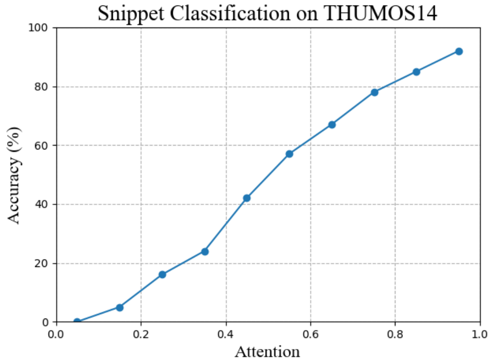

Figure 1. Visualization for accuracy curve of the base model. Classification accuracy for $\mathbf{Z}$. Over our threshold τ, the accuracy is over 80%.

For L_{att}, we take a threshold to preserve the snippets with higher classification accuracy. In Fig. 1, we show the snippet classification accuracy of $\boldsymbol{Z}$ by attention value. It is evident that higher attention values correlate with improved classification accuracy. We retain snippets with attention values exceeding $\tau = 0.8$, ensuring that the regression-based model learns from samples with over 80% accuracy.

References

[1] Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah, and Mei Chen. Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22992–23002, 2023. 1

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

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

[4] Jingqiu Zhou, Linjiang Huang, Liang Wang, Si Liu, and Hongsheng Li. Improving weakly supervised temporal action localization by bridging train-test gap in pseudo labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 23003–23012, 2023. 1