Realigning Confidence with Temporal Saliency Information for Point-Level Weakly-Supervised Temporal Action Localization

Supplementary Material

Table 10. Ablation results of different updating thresholds θ_{up} for the saliency point updating on THUMOS'14.

Setup	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
$\theta_{up} = 0.6$	81.7	69.1	48.7	23.7	56.7
$\theta_{up} = 0.7$	82.3	69.7	49.4	23.7	57.1
$\theta_{up} = 0.8$	**82.3**	**70.1**	**49.4**	**24.5**	**57.4**
$\theta_{up} = 0.9$	82.1	69.8	49.1	23.3	57.0

Table 11. Ablation results of different updating periods p_{up} for the saliency point updating on THUMOS'14.

setup	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
$p_{up} = 50$	82.1	69.3	49.2	23.2	56.8
$p_{up} = 100$	81.9	69.4	49.5	24.2	57.0
$p_{up} = 200$	**82.3**	**70.1**	**49.4**	**24.5**	**57.4**
$p_{up} = 400$	82.0	69.6	49.2	24.1	57.0

Table 12. Effect of the related information range for ABA on THUMOS'14.

Setup	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
$\theta_{re} = 0.2$	81.7	69.4	48.4	23.0	56.5
$\theta_{re} = 0.4$	**82.3**	**70.1**	**49.4**	24.5	**57.4**
$\theta_{re} = 0.6$	81.7	69.8	48.4	**24.7**	56.9
$\theta_{re} = 0.8$	81.5	69.5	48.2	23.6	56.6

Table 13. Impact of point selection on performance in proposals containing multiple saliency points.

Setup	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
first	**82.3**	**70.1**	**49.4**	24.5	**57.4**
last	82.1	69.6	49.2	24.4	57.2
center	81.7	69.0	48.5	23.9	56.6
average	81.9	69.6	48.6	**24.6**	56.9
soft-value	82.0	69.4	48.6	23.7	56.8

7. Supplementary Details

Tendency of annotator. The primary motivation of this paper stems from the observation that annotators consistently tend to label the most salient frame as the point label for each instance. This inherent behavior inspires us to establish a relationship between the single-frame labels and the proposals' quality. In Figure 2, we utilize the labels provided in SF-Net [33] to analyze the frequency of point locations within instances across three benchmarks. The statistical results reveal that the majority of points are situated in the central region, with annotators rarely labeling the boundaries. We posit that this trend arises from the inherent difficulty of recognizing the category of each instance at the ambiguous beginning region. Additionally, it is often impractical to watch the entire instance and still fail to recognize it. Thus, this centered labeling bias holds significant value in supervising the completeness of proposals, a point that we demonstrate in the main text.

Proposal Generation. The approach to proposal generation significantly influences the effectiveness of proposal learning and the ultimate quality of localization. P-MIL [39] suggests a valid strategy of generating both background and action proposals. However, we observed no improvement, and in some cases, a decrease in performance when applying this strategy to our proposed method. Upon further analysis, we identified that the primary function of generating background proposals in P-TAL is to balance the ratio of positive and negative samples. Given the presence of point labels in P-TAL, the generated proposals can be naturally divided into positive and negative samples. Consequently, after compiling the statistics of positive and negative samples, we consider background proposal generation as an alternative. Specifically, we implement background generation on GTEA and BEOID at the thresholds $\theta_{bg} = 0.1, 0.3$, considering that daily videos in these datasets exhibit continuous and dense action instances.

8. Additional Ablations

The scope and period of updating saliency points. For updating the saliency points, the update threshold θ_{up} determines the number of proposals to be used for updating, and the update period p_{up} affects the stability of saliency points. As shown in Table 10, the small θ_{up} will introduce the updating interference, and the too-large one will result in less effective updates as little of the proposal information is used. In Table 11, the performance exhibits stability at the long update periods, which proves the human nature that most of the annotated points are salient. However, the small p_{up} will cause turbulence for the saliency points.

The related information range for boundary adap-

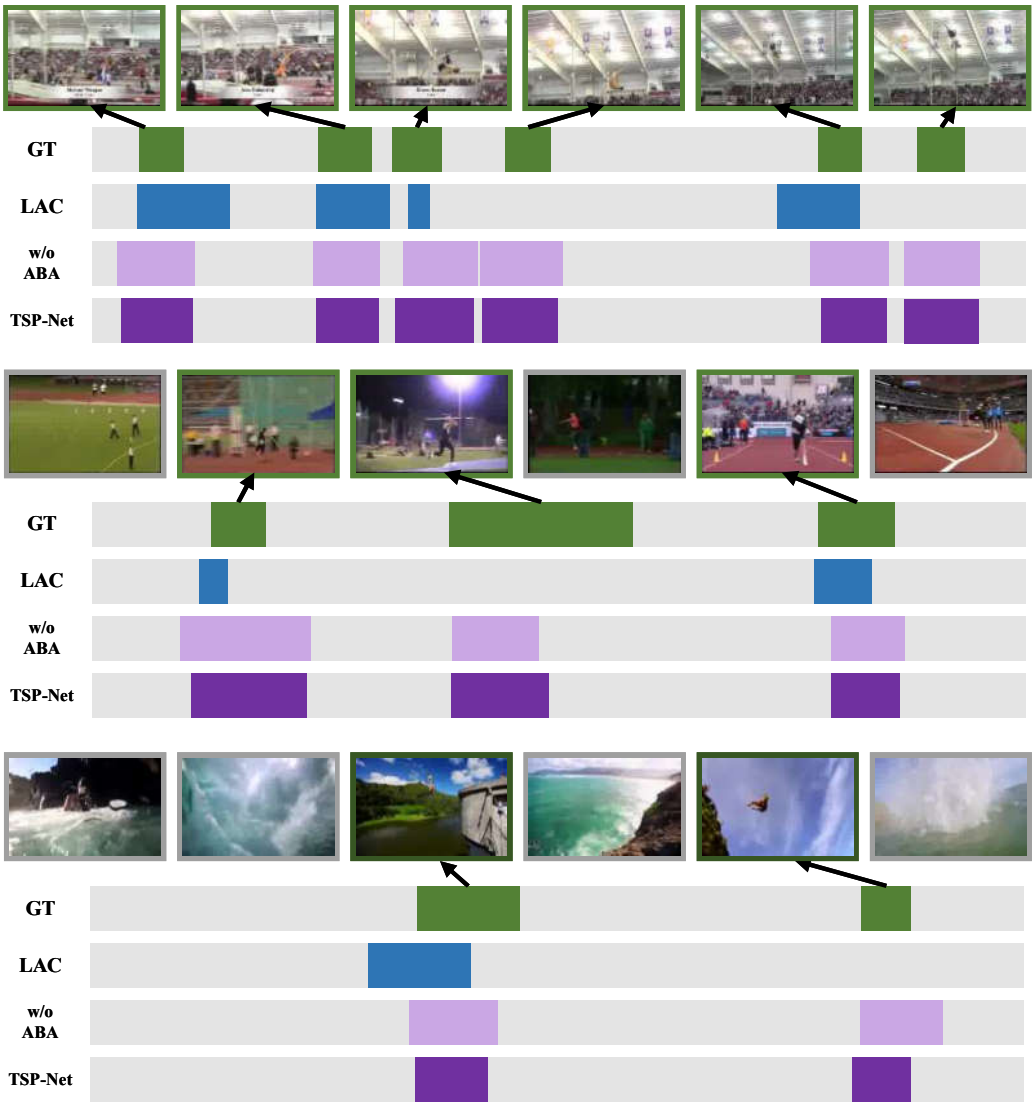

Figure 6. Addition visualization of the localization of the base LAC and the proposed method that applies ABA or not (IoU > 0.3).

tion is controlled by θ_{re}. We conduct an ablation study to explore the characteristic in Table 12. Similar to the updating thresholds in Table 10, the extreme θ_{re} will introduce noisy or useless boundary information and cause sub-optimal adaption. However, the best performance is achieved when θ_{re} is set to 0.4. The results indicate that compared with using proposals to update saliency points, the information of low-quality proposals is restrained under the premise of obtaining aligned confidence. On the other hand, based on the IoU information, the proposals with low IoU used for adaption will be suppressed due to their temporal irrelevance. Thus, more proposal information can effectively strengthen the proposal's completeness.

Proposals containing multiple points. For the computation of center label generation in Eq. 2 and Eq. 4, the ideal situation is that each proposal contains only one saliency point or none. However, our observations reveal the presence of a small number of proposals that contain multiple salient points (approximately 1/10 of the total). These proposals consistently exhibit long durations and are of low quality. Therefore, the selection of salient points has a significant impact on the quality of generated saliency labels. To investigate this situation's influence on localization, we established five different selection methods, as

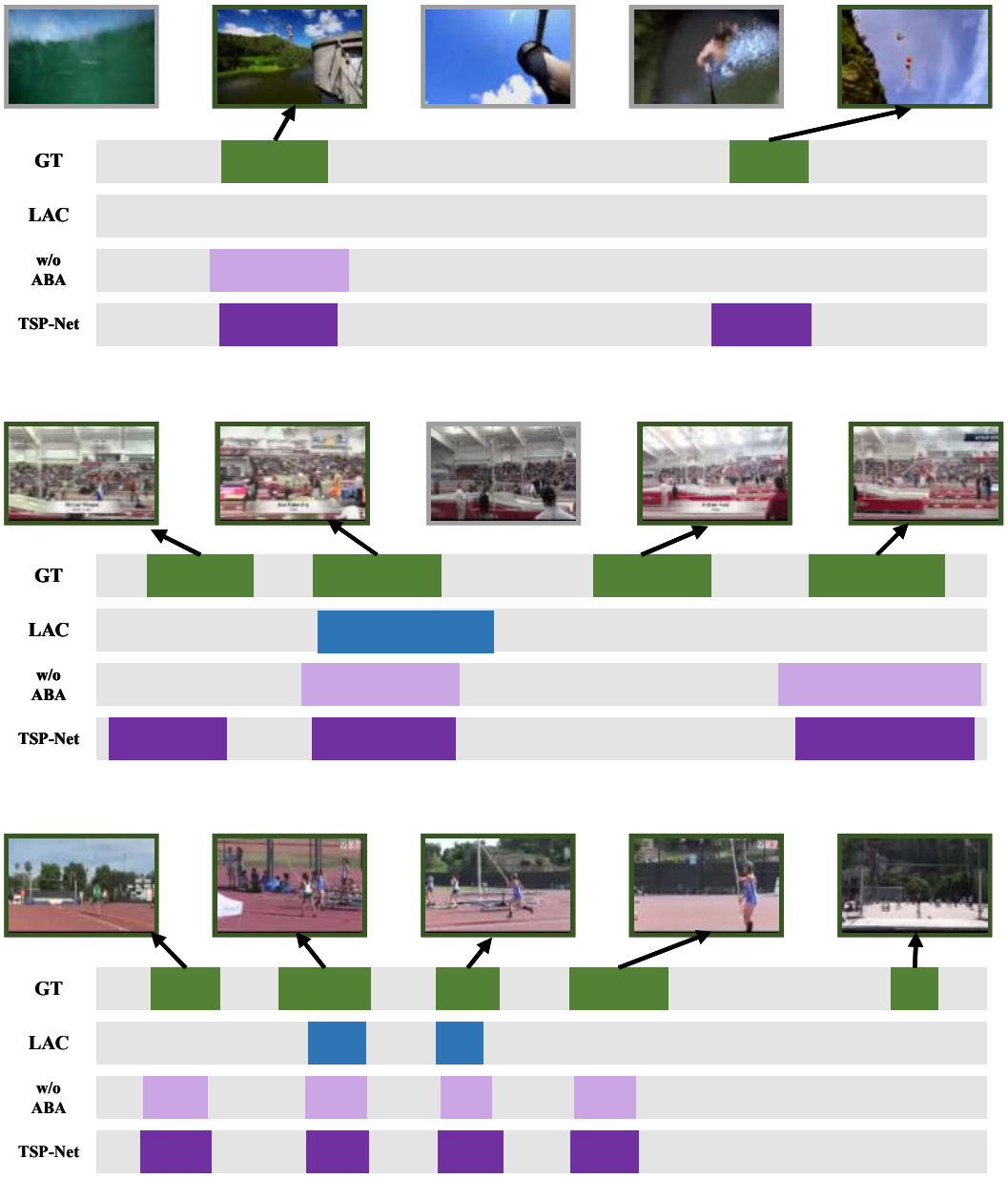

Figure 7. Additional visualization of the localization of the base LAC and the proposed method that applies ABA or not (IoU > 0.6). The proposed ABA adaptively utilizes valuable boundary information to achieve localization with high completeness.

shown in Table 13. Experimental results indicate that using the first saliency point to generate the saliency label yields the best localization performance. This is because the first point tends to be located at the early position of the proposal, aligning well with both those proposal quality and the generated label. To further validate this point, we also generated labels using the last point and the center point, respectively. The center scores learned from the last point-generated labels achieve competitive performance compared to the first ones, suggesting that they share the same generative idea, where the localization results are solely determined by the location distribution of the annotated points. In contrast, selecting the center point yields the worst results, as it tends to align with the proposal cen-

ter, resulting in unreliable center labels. The average setting performs better than the central setting, demonstrating that calculating the average of multi-point generated labels can alleviate the aforementioned drawbacks. Lastly, simply manually assigning soft values to the center labels of proposals with multiple points does not yield effective results, as it disregards the alignment relationship between proposals and complete action instances.

9. Additional Qualitative Analyses

We show more representative qualitative localization results at different IoUs in Figure 6 and Figure 7 to prove the superiority of our proposed method, respectively. Specifically, as mentioned in the main text, our method leverages learned aligned confidence to improve the matching of correct proposals with real ground-truth instances, particularly when the IoU is low (IoU=0.3). This characteristic allows our method to achieve a significant improvement in mAP at low IoUs compared to the baseline. In Figure 7, we provide visualizations of the localization results under the constraint that the IoU is greater than 0.6. It can be observed that most baseline detection results fail to achieve accurate localization when completeness is expected. Furthermore, TSP-Net also struggles to achieve effective localization without utilizing the proposed alignment-based boundary adaption. This is because the original proposal generation strategy is overly sensitive, resulting in unreliable completeness of the generated proposals, which cannot be mitigated even with the use of aligned confidence. However, with the introduction of the ABA strategy, we adapt the proposal boundaries by considering close-set information and the correlation between proposals. By optionally incorporating boundary information, we can alleviate the impact of the aforementioned proposal generation characteristic on localization completeness.

10. Failure Analyses and Future Work

Although our proposed TSP-Net achieves effective detection performance improvement compared with the baseline, it still achieves unsatisfactory positioning results for some instances. In particular, in Figure 7, TSP-Net has missed detection when the viewpoint changes significantly. The above theoretical and qualitative analyses lead to our future work, including but not limited to (1) Generating more discriminative center labels based on the original information. (2) Designing a two-stage localization network based on the temporal saliency information to maintain semantic consistency. (3) Mining more precise temporal saliency information from different types of proposal.