# *Supplementary Material*: Proposal-Free Temporal Action Detection via Global Segmentation Mask Learning

Sauradip Nag[1,2], Xiatian Zhu[1,3], Yi-Zhe Song[1,2], and Tao Xiang[1,2]

[1] CVSSP, University of Surrey, UK
[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence, UK
[3] Surrey Institute for People-Centred Artificial Intelligence, University of Surrey, UK
{s.nag,xiatian.zhu,yz.song,t.xiang}@surrey.ac.uk

## 1 Appendix A: More Implementation Details

**A.1 More details on boundary IoU loss** Intersection over Union (IoU) is the standard evaluation metric for segmentation (*e.g.*, image segmentation) and detection tasks (*e.g.*, object detection and temporal action detection). Given a number of predictions it measures what are true positives and false positives against the ground-truth. Besides, IoU has been successfully used to design loss functions [15, 9] for training object detection models. Recently, it is shown that boundary region is critical part for mask prediction in image segmentation.

Inspired by these considerations as above, in this work a novel boundary IOU (bIoU) loss design is introduced. We first define the boundary IOU metric. Concretely, given a ground-truth mask $G$ (Fig. 2(a)), we obtain the boundary region $G_d$ by extracting those pixels within a given distance $d$ away from the contour (Fig. 2(c)). This can be implemented using a morphological erosion operation, with the parameter $d$ controlled by the morphological kernel $k$. Given a predicted mask $P$ (Fig. 2(b)) we similarly obtain the boundary region $P_d$ (Fig. 2(d)). We then compute the bIoU metric between $G$ and $P$ as:

$$bIOU(P,G) = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|} \tag{1}$$

The distance parameter $d$ controls the boundary sensitivity. Note, when $d$ is sufficiently large, the boundary will expand to the whole mask, leading to the standard IoU metric. Empirical evaluations in Table 1 suggests that selecting $k = 7$ gives the best temporal boundaries.

We formulate the bIoU loss function as:

$$\begin{aligned} L_{bIOU} = 1 - bIOU(P,G) &= 1 - \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|} \\ &= 1 - \frac{|\Phi(G) \cap \Phi(P)|}{|\Phi(G) \cup \Phi(P)|} \end{aligned} \tag{2}$$

where $\Phi(.)$ represents the morphological operation.

Table 1: Impact of kernel size on bIOU loss on ActivityNet.

| Kernel Size | mAP | |
|---|---|---|
| | 0.5 | Avg |
| 3 | 53.3 | 34.3 |
| 5 | 55.1 | 35.7 |
| 7 | **56.3** | **36.5** |
| 9 | 55.7 | 36.0 |

**Differentiation** To facilitate model training, a loss function needs to be differentiable. To that end, we adopt the differentiable morphological erosion [10] as the mask morphological operation.

**Vanishing gradients** Our proposed bIOU loss may suffer from vanishing gradients, causing extra convergence difficulties. For example, this may arise at the cases of non-overlapping mask boundaries, as illustrated in Fig. 2(d,f). To alleviate this problem, we append an additional $L_2$ distance penalty on the whole mask. Mathematically, the entire bIoU loss can be finally defined as:

$$\begin{aligned} L_{bIOU} &= 1 - bIOU(P,G) + \lambda * L_2(P,G), \\ &\text{with } \lambda = 1/(\Phi(G) \cap \Phi(P) + \epsilon). \end{aligned} \tag{3}$$

Here, $\lambda$ specifies the coefficient inversely proportional to the boundary overlap and $\epsilon$ is used for avoiding zero denominator. We conjecture that this penalty can help push the predicted boundaries closer to the ground-truth to yield better masks. Concretely, this penalty term is designed as:

$$L_2(P,G) = \frac{d(P,G)}{c} \tag{4}$$

where $d(.)$ is the normalized $L_2$ distance between the masks $P$ and $G$, and $c$ is the foreground snippet number of ground-truth mask. We found that using this additional penalizing term, the model training can became more stable and better fit to the training data (Fig. 1).

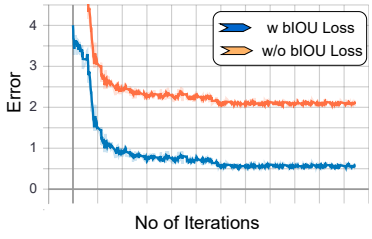


Fig. 1: Convergence of the mask branch's loss.

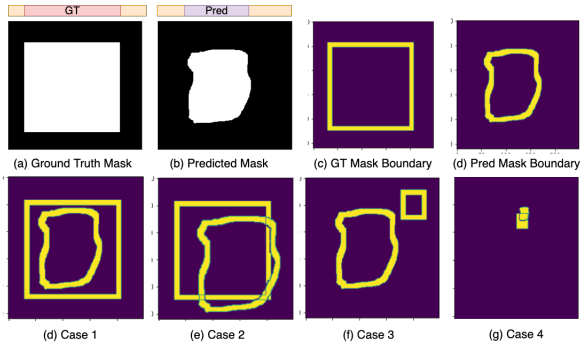


Fig. 2: A 2-D illustration of our bIOU loss design and example cases.

**Comparing IoU loss and bIoU loss on Temporal Action Detection** For quantitative evaluation, we compare the IOU loss $L_{IOU}$ and our proposed bIOU loss $L_{bIOU}$ for training the masl branch of TAGS. The dice loss $L_{dice}$ is also applied simultaneously. Table 2 shows that our bIOU loss gives a performance gain of 1.1% in average mAP on ActivityNet, due to its ability of predicting better temporal boundaries.

Table 2: Analysis of IOU and bIOU loss on ActivityNet.

| Loss | mAP | |
|---|---|---|
| | 0.5 | Avg |
| $L_{dice} + L_{IOU}$ | 54.9 | 35.3 |
| $L_{dice} + L_{bIOU}$ | **56.3** | **36.5** |

**IoU loss vs. bIoU loss on RGB saliency detection** Image saliency is essentially a mask prediction task. In this test, we use a popular saliency detection method BASNet [8] on DUTS dataset [11]. The performance metrics are Mean Absolute Error (MAE) and F-measure of boundary ($F_\beta$). To train the model, binary cross-entropy (BCE) and SSIM losses are also applied in addition to IoU/bIoU loss. Table 3 shows that our bIOU loss is again superior to the conventional IoU loss, suggesting its general advantage.

Table 3: Comparative analysis of IOU and bIOU loss on image saliency detection. BASNet [8] is used.

| Loss Head | DUTS Dataset | | |
|---|---|---|---|
| | $maxF_\beta$ | $relaxF^b_\beta$ | MAE |
| $L_{bce} + L_{ssim} + L_{iou}$ | 0.942 | 0.826 | 0.037 |
| $L_{bce} + L_{ssim} + L_{biou}$ | **0.940** | **0.823** | **0.034** |

## 2 Appendix B: More Analysis

**Importance of structured consistency** We further analyze the impact of structural consistency on the fine-grained performance by video length. Following [1], the videos of THUMOS dataset are classified into 5 different categories by action temporal duration: extra-small, small, medium, long and extra-long. We compare our TAGS with and without the structural consistency: (i) mask redundancy ($L_{pp}$) or (ii) class-mask consistency ($L_{fc}$) on each of these 5 duration categories individually. As seen in Fig. 3, our TAGS performs better with structural consistency especially on more challenging short videos with less action content.

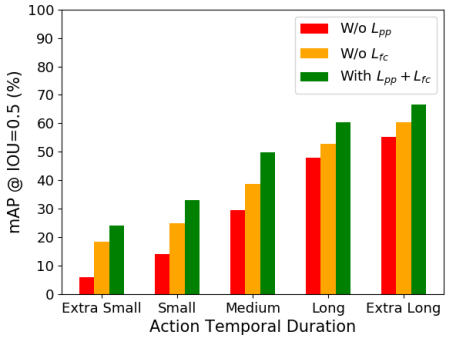


Fig. 3: **Impact of structural consistency across different video lengths on THUMOS dataset.**

**Scoring method** We evaluate the design of sequence scoring (refer to Eq. (7) of main paper) on ActivityNet. We investigate three distinct scoring methods for training: (a) only inner scores, (b) score contrast of action instances, (c) contrast of both action and background ones, which is our default design. As shown in Table 4, compared to inner scores, both contrast methods generate more accurate mask sequences and hence bring good performance gains at high IoU thresholds. Moreover, incorporating background segments in score calculation helps to find more accurate mask sequences thereby improving the detection performance at test time.

**TAGS with conventional objective loss** To evaluate the importance of our loss functions, we test TAGS under common conventional TAD loss functions on ActivityNet. We evaluate the performance of our TAGS with simple loss functions: the wBCE loss used in BMN [2] for the mask branch and the standard cross entropy loss for the classification branch. We denote this variant as TAGS$^{\dagger}$. For fair comparison, we use BMN [2] as a competitor using the same post-processing and the feature backbone. It can be observed in Table 5 that the

Table 4: Ablation study on choice of mask-redundancy score on ActivityNet.

| Scoring design | mAP | | | |
|---|---|---|---|---|
| | 0.5 | 0.75 | 0.95 | Avg |
| Inner Score | 54.9 | 35.2 | 9.1 | 35.3 |
| Contrast in Action | 55.5 | 36.0 | 9.3 | 35.7 |
| **Contrast in Action and Background** | **56.3** | **36.8** | **9.6** | **36.5** |

performance of our single stage TAGS$^{\dagger}$ drops by 4.2% in mAP@0.5, justifying the importance of our loss design (refer to Table 6 in main paper). Further, with even such a simple loss function our TAGS can still be comparable to BMN [2], justifying our model design.

Table 5: Ablation of TAGS with conventional objective loss.

| Method | mAP | |
|---|---|---|
| | 0.5 | Avg |
| **TAGS (ours)** | **56.3** | **36.5** |
| TAGS$^{\dagger}$ | 52.1 | 33.8 |
| BMN [2] + UNet [12] | 50.1 | 33.9 |

**Analysis of component design in TAGS** Our TAGS primarily consists of a Snippet Embedding Transformer and 1-D Convolution heads for classification and localization branch. We ablate the number of 1-D CNN layers for both the branch heads in Table 6. As the results suggest, only 1 layer is enough for classification branch. A plausible reason for this is that for classification it needs global information and stacking multiple 1-D CNN may affect global information. For localization branch, it is observed that 3 layers give best performance. This is probably because for predicting the masks the network needs to process local information captured by 1-D CNNs. Additionally, we also ablate the performance of transformer design in head size. Table 7 demonstrates that the performance of TAGS improves significantly with the increase of heads in the Transformer. However, excessive heads will lead to overfitting. The performance peaks at four heads.

**Cross-domain generalization** The experiments so far assumed that the training and test data come from the same dataset/domain. However, in real-world applications a trained model typically needs to handle many different deployment situations out of the box. To simulate this more realistic deployment setting, we design a cross-domain experiment using a subset of classes shared by ActivityNet and THUMOS. We manually match the class semantics across the two datasets and then merge those classes with same semantics but different names. This results in a total of 12 classes. G-TAD [14] is selected for comparative evaluation. We then train each model on one dataset and test on the other. We observe from Table 8 that: (1) Both models' performance degrades (vs. Table 1

Table 6: Effect of the number of 1-D CNN Layers for the classification and mask branches on ActivityNet.

| # Layers | Class. brch | | Mask brch | |
|---|---|---|---|---|
| | 0.5 | Avg | 0.5 | Avg |
| 1 | **56.3** | **36.5** | 52.7 | 34.2 |
| 2 | 55.8 | 36.0 | 53.8 | 35.1 |
| 3 | 55.2 | 35.9 | **56.3** | **36.5** |
| 4 | 54.3 | 35.1 | 56.0 | 36.1 |
| 5 | 53.8 | 34.7 | 55.9 | 36.0 |

Table 7: Impact of the head number in the Transformer on ActivityNet.

| Number of heads | mAP | |
|---|---|---|
| | 0.5 | Avg |
| 1 | 53.8 | 34.8 |
| 2 | 54.6 | 35.0 |
| 3 | 55.2 | 35.7 |
| 4 | **56.3** | **36.5** |
| 5 | 55.8 | 35.9 |

(main)) under this more challenging setting due to the data distribution shift. (2) Importantly, our model's advantage over G-TAD is even bigger compared to the same-domain setting, suggesting that our model is more suited to real-world deployments. This is not surprising as simpler models often generalize better.

Table 8: Cross-domain generalization.

| **Methods** | **ActivityNet → Thumos** | | **Thumos → ActivityNet** | |
|---|---|---|---|---|
| | **mAP@0.5** | **Avg mAP** | **mAP@0.5** | **Avg mAP** |
| GTAD [14] | 27.5 | 28.2 | 34.5 | 22.1 |
| **TAGS** | **32.7** | **30.3** | **43.4** | **25.6** |

**Transformers in existing TAD setting** We examine how well existing TAD methods [4, 5, 7, 6, 13] work with TAGS's transformer for snippet embedding. We select a representative model BMN [2] and insert our snippet embedding module right after the video encoder. As shown in Table 9, self-attention can also improve the performance of BMN, demonstrating the importance of temporal relationship modeling for temporal action detection task. However, it is still significantly inferior to our TAGS model.

Table 9: Transformer for existing TAD methods on ActivityNet.

| **Network** | **mAP** | |
|---|---|---|
| | **0.5** | **Avg** |
| BMN [2] | 50.1 | 33.9 |
| Transformer + BMN | **51.6** | **34.8** |

**Effect of snippet length** We evaluate the impact of video snippet length $T$ for TAGS on ActivityNet. As shown in Table 10, when the snippet length is small (*e.g.*, 100), the pooling of multi-scale will further bring down to extremely small temporal snippet dimension. As a result, we observe a performance drop of 2.8% in mAP@0.5. This is as expected in that too few snippets per video are less capable to represent local motion patterns. We find that once the snippet length increases until 400, the performance drop keeps reducing. Further increasing the temporal dimension to 800 points gives the best result in terms of cost. In this case, we apply 3 temporal scales of 100, 200, 400 snippets throughout, and impose the training supervision to the top two scales (100 and 200). Further including the 400 scale does not give benefit, as shown in Table 5 of main paper. Further increasing the snippet dimension indeed boosts the performance slightly, whilst at a more significant computational cost.

Table 10: Impact of the snippet length of a video on ActivityNet.

| Snippet Length | mAP | |
|---|---|---|
| | 0.5 | Avg |
| 100 | 53.5 | 34.8 |
| 200 | 54.9 | 35.2 |
| 400 | 55.7 | 36.1 |
| **800** | **56.3** | **36.5** |
| 1000 | 56.5 | 36.6 |

**Effect of loss objectives** The results in Table 11 show that each loss is beneficial for TAD's accuracy. In particular, focal loss can tackle the class imbalance problem in classification branch with 4.5% gain in Avg mAP, balanced logistic regression (LR) loss (Eq. 5 in main paper) treats the snippet classes individually by taking a binary mask problem, binary dice loss (Eq. 6 in main paper) handles the imbalance problem between action and background classes, whilst our proposed boundary IOU (bIOU) loss (Eq. 6 in main paper) is helpful in sharpening the foreground mask prediction. More specifically, bIOU contributes 3.1% in mAP@0.5 and 2.6% in Avg mAP, indicating the importance of temporal boundary and the effectiveness of our loss design in regulating more capacity for boundary inference. Besides branch-specific loss terms, the mask redundancy

loss term $L_{pp}$ benefits in IOU@0.5 and specially in stricter IOU scores. This reveals the reason for low localization error in proposal-free designs like TAGS [5, 4, 6, 7]. The classification-mask consistency loss $L_{fc}$ also contributes 0.5% in mAP@0.5 and 0.3% in avg mAP, validating the usefulness of structural consistency in design.

Table 11: Effect of TAGS loss objectives on ActivityNet.

| **Loss** | **mAP** | |
|---|---|---|
| | **0.5** | **Avg** |
| TAGS (full) | 56.3 | 36.5 |
| w/o Focal Loss | 51.4 | 32.0 |
| w/o Balanced LR Loss | 55.2 | 35.4 |
| w/o bIOU Loss | 53.2 | 33.9 |
| w/o Dice Loss | 52.5 | 32.7 |
| w/o Mask Red. Loss ($L_{pp}$) | 55.6 | 36.0 |
| w/o Const. Loss ($L_{fc}$) | 55.8 | 36.2 |

**Effects of action instance duration** We additionally evaluate how the model performance is affected by the duration of action instances on THUMOS. We compare our proposal-free method with a proposal based approach BMN [2]. We measure segmentation error between the ground-truth and temporal prediction both with $L_1$ norm) against the ground-truth normalized duration. As seen in Fig. 4, our TAGS [5] yields lower segmentation error than BMN particularly for shorter action instances w.r.t. the whole video length.

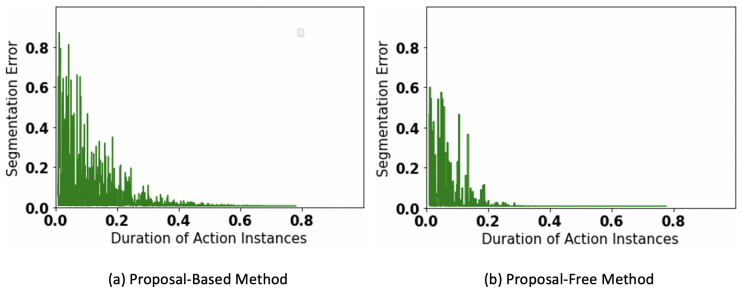


Fig. 4: Segmentation error analysis on THUMOS.

**Inference speed** For comparison with more previous methods with no training code released, we can only compare with their reported inference speed measured in FPS without considering the feature extraction time similar to [3]. As different GPU hardware is used in previous papers, for easier comparison we translate the

FPS speed according to their specification. For this comparison, I3D features on THUMOS14 are used. It is evident from Fig. 5 that despite being a multi-scale network, our TAGS runs much faster, *e.g.*, 3/4× faster than PGCN/SSTAD. This is because our TAD model is light-weight with only a light Transformer and several 1-D conv blocks.

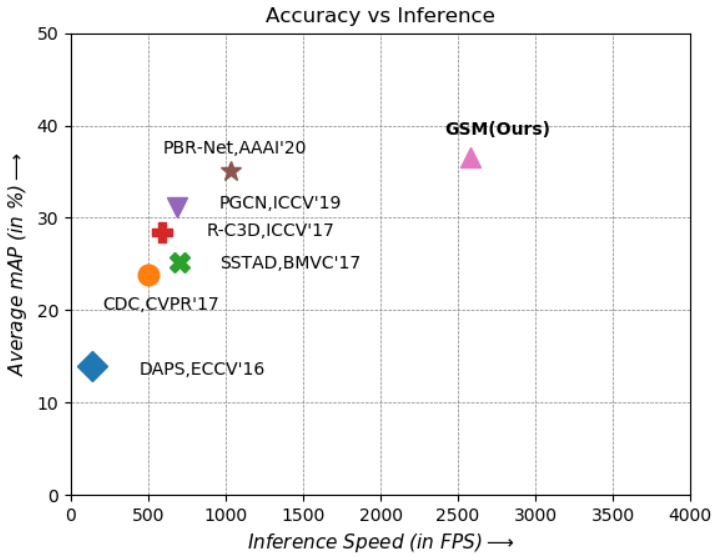


Fig. 5: Accuracy vs. speed (translated FPS based on Titan XM).

## 3 Qualitative Results

In this section, to make more visual examination we provide additional qualitative results by GTAD [14] and our TAGS model on both ActivityNet and THUMOS dataset. We focus on two challenging situations: (i) a single short-duration action instance per video (Fig. 7), and (ii) multiple short-duration action instances per video (Fig. 8). From these examples, we have a similar observation that compared to G-TAD, our proposed TAGS method can localize the target action instances more accurately with a much smaller number of outputs, thus being more efficient at inference.

## References

1. Alwassel, H., Heilbron, F.C., Escorcia, V., Ghanem, B.: Diagnosing error in temporal action detectors. In: ECCV. pp. 256–272 (2018)

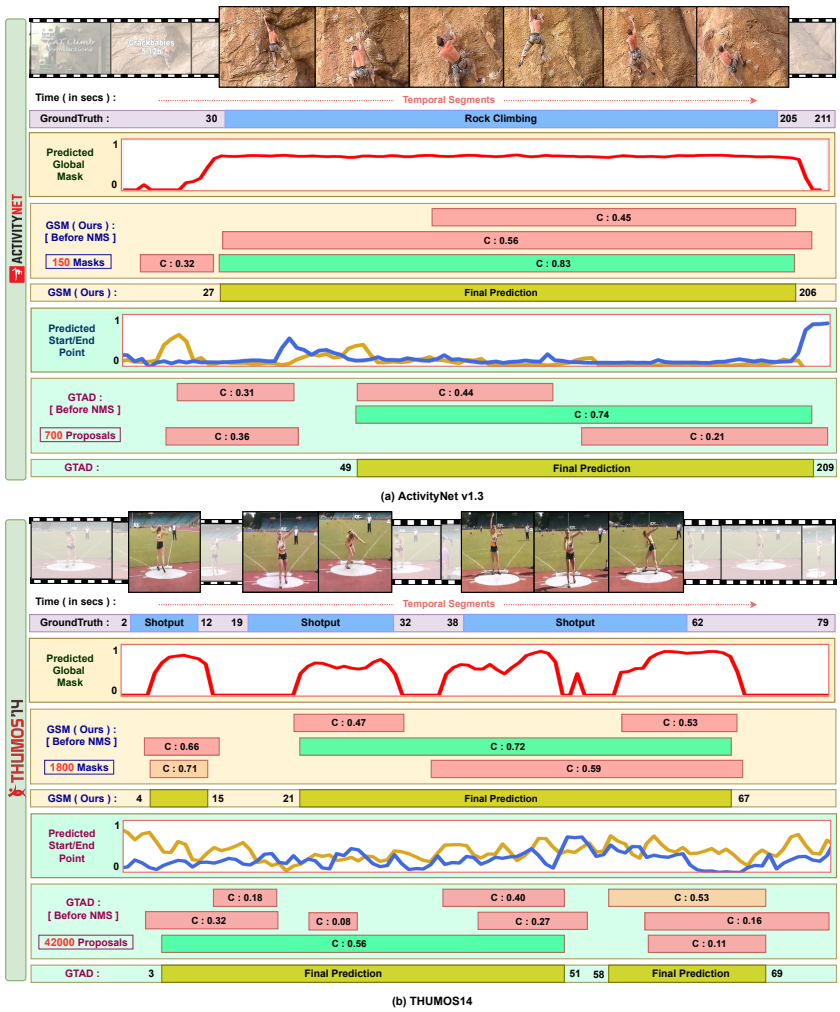


Fig. 6: **Qualitative TAD result comparison** on videos from (a) ActivityNet-v1.3 and (b) Thumos14. We compare our TAGS (first 3 rows) with G-TAD [14] (last 3 rows). For each method, we show a number of top action detection candidates, with the confidence score given inside each detection box. It can be seen that for both cases, our TAGS produces more accurate action instance detection with much less candidates compared to G-TAD.

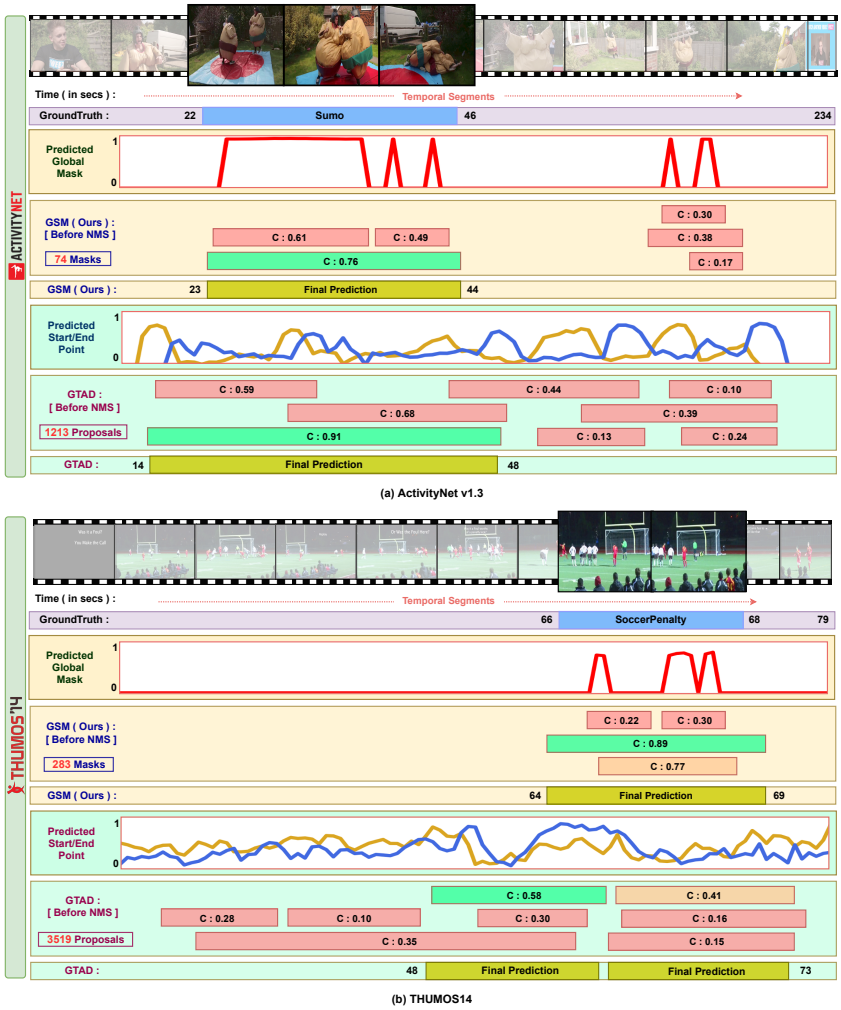


Fig. 7: **Qualitative TAD result comparison** on single-instance videos from (a) ActivityNet-v1.3 and (b) Thumos14.

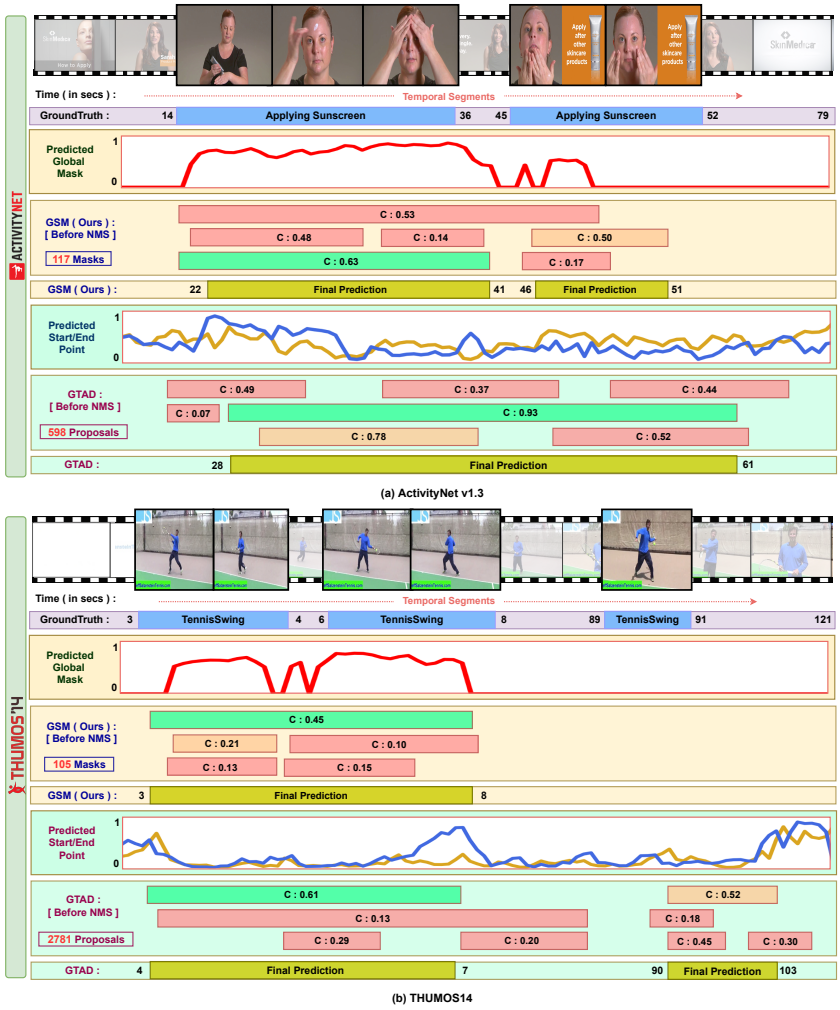


Fig. 8: **Qualitative TAD result comparison** on multi-instance videos from (a) ActivityNet-v1.3 and (b) Thumos14.

2. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: Bmn: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3889–3898 (2019)
3. Liu, Q., Wang, Z.: Progressive boundary refinement network for temporal action detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11612–11619 (2020)
4. Nag, S., Zhu, X., Song, Y.Z., Xiang, T.: Temporal action localization with global segmentation mask transformers (2021)
5. Nag, S., Zhu, X., Song, Y.z., Xiang, T.: Proposal-free temporal action detection via global segmentation mask learning. In: ECCV (2022)
6. Nag, S., Zhu, X., Song, Y.z., Xiang, T.: Semi-supervised temporal action detection with proposal-free masking. In: ECCV (2022)
7. Nag, S., Zhu, X., Song, Y.z., Xiang, T.: Zero-shot temporal action detection via vision-language prompting. In: ECCV (2022)
8. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7479–7489 (2019)
9. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union (June 2019)
10. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: an open source differentiable computer vision library for pytorch. In: WACV. pp. 3674–3683 (2020)
11. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017)
12. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: CVPR. pp. 4325–4334 (2017)
13. Xu, M., Pérez-Rúa, J.M., Escorcia, V., Martinez, B., Zhu, X., Zhang, L., Ghanem, B., Xiang, T.: Boundary-sensitive pre-training for temporal localization in videos. arXiv (2020)
14. Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B.: G-tad: Sub-graph localization for temporal action detection. In: CVPR (2020)
15. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: ACM MM. pp. 516–520 (2016)