Supplementary Material for Learning Salient Boundary Feature for Anchor-free Temporal Action Localization

Chuming Lin*[1], Chengming Xu*[2], Donghao Luo[1], Yabiao Wang[1], Ying Tai[1],
Chengjie Wang[1], Jilin Li[1], Feiyue Huang[1], Yanwei Fu[2]
[1]Youtu Lab, Tencent, [2]Fudan University, China
{chuminglin, michaelluo, caseywang, yingtai, jasoncjwang, jerolinli, garyhuang}@tencent.com
{cmxu18, yanweifu}@fudan.edu.cn

1. Experimental detail

1.1. Detailed Architecture for Feature Extractor

We plot the detailed architecture of our feature extractor in Fig. 1. A pre-trained I3D [1] is used to process the video X, with which the outputs of two levels named *C4* and *C5* are then used in the FPN. We utilize a 6 layer FPN. These six feature maps are further processed with our basic predictor and saliency-based refinement module for temporal localization.

1.2. Training Algorithm

The whole training algorithm is shown in Alg. 1, where both ordinary objective functions for TAL and our novel Boundary Consistency Learning are employed to optimize our model.

Algorithm 1 Training Anchor-Free Saliency-based Detector

Require: $\mathcal{T}^{train}$: distribution of training dataset
while not converge **do**
  Sample batch of tasks $B \sim \mathcal{T}$
  for all FPN layer l **do**
    for all temporal location i **do**
      Get all coarse and refined predictions $\hat{\psi}_{l,i}, \hat{\xi}_{l,i}, \hat{y}^C_{l,i}, \Delta\hat{\psi}_{l,i}, \Delta\hat{\xi}_{l,i}, \hat{y}^R_{l,i}, \eta_{l,i}$
    end for
  end for
  Calculate objective function $\mathcal{L}$ as in Eq. 11
  Update model parameters according to $\mathcal{L}$
  Batch for Boundary Consistency Learning $\hat{B} = \{\}$
  for all video X_i in B **do**
    Calculate minimum action length w_{min} in X_i
    if X has an action instance whose length is larger than $2w_{min}$ and a background clip with length w_{min}, **then**
      $\hat{B} = \hat{B} \bigcup X_i$
    end if
  end for
  for all video X_i in $\hat{B}$ **do**
    Calculate ℓ^i_{con} as in Eq. 10.
  end for
  Update model parameters according to $\ell_{con} = \frac{1}{|\hat{B}|}\sum_i \ell^i_{con}$
end while

2. Additional Experiment Results

2.1. Verification of Dynamic FPS

In former work [2], the video is extracted frames via using fixed frame per second (fps) and sliding window to reduce the input frame number of the model. This strategy is suitable for THUMOS14, since the duration of most action instances is short and in range $(0s, 20s]$. We can use a common fps value, such as 10 fps in our work, to sample frames for each video. Then each video is split into clips with 256 frames using sliding windows. Therefore, each clip can cover 99.7% ground truths whose duration is less than 25.6 seconds.

However, on ActivityNet1.3, the duration range of action instances is wide and most videos contain a very long action instance which almost cover the video. Thus, the input clip should include the full video to ensure the model can predict a complete action, and sliding window is not a good choice to this dataset which will split the ground truth. But, we can utilize a small and fixed fps such as 3 fps to sample frames and set clip length to 768 for ActivityNet1.3, because the sample frame number of 99.9% videos is less than 768 frames. Nevertheless, extracting frames using a small fps will drop out more video information. To solve

* indicates equal contributions.
This work was done when Chengming Xu was an intern at Tencent Youtu Lab. Yanwei Fu is the corresponding author.

this problem, we propose a new sample approach, named *Dynamic FPS*, to calculate exclusive sample fps and make sure the sample frame number is 768 for each video. For example, given a video with 96 seconds, the sample fps is set to 8 to extract 768 frames. As shown in Tab. 1, Compared with sampling with a small fixed fps, the Dynamic FPS sampling method improves 1.7% average mAP on ActivityNet1.3 that means the strategy is able to retain as much information as possible in same clip length.

2.2. Ablation Study on Flow and Fusion Model

Apart from the ablation study in our main paper conducted on RGB model, we further provide a brief ablation study on both flow and fusion model in Tab. 3. This experiment looks into these parts of our model: (1) the Group Normalization, (2) the quality term η and (3) our refinement strategy. We find that: (1) Our anchor-free method can have comparable performance with anchor-based ones in a fair setting without GN and Boundary Consistency Learning. (2) Adding other components can lead to further improvement, which is consistent with the analysis achieved by rgb-only model. (3) Results on ActivityNet are consistent to those on THUMOS14.

2.3. Ablation Study on Soft-NMS

For further comparison we experiment with a variant of our model with NMS. THe results on THUMOS14 are shown in Tab. 2. The model with NMS has a 1.6 average mAP gap with that using soft-NMS. However, the performance is still better than that of G-TAD+PGCN, indicating the efficacy of our method.

2.4. More Visualization Results

Apart from quantitative results in the main paper, we visualize the two shots of our model along with anchor-based R-C3D on an action instance. Fig. 3(a) shows that the refinement procedure improves a rough prediction generated by the naive predictor to be better than the prediction of R-C3D. We take the channel-wise mean of each half and plot it in Fig. 3(b), where we can see that when trained with BCL, the 'start' feature would have a peak when the action starts, the same for the 'end' feature, while the other regions of background and internal action keep a low activation, which reflects that we successfully learn both start-sensitive and end-sensitive features. This also supports the above analysis that only with the proposed Boundary Consistency Learning, the boundary pooling can have reasonable features to be processed with, thus having good results.

Moreover, we visualize more predictions produced by R-C3D [2] and both coarse and refined stages of our model in Fig. 3. The results are still consistent and supports our claim that while our coarse predictor can already provide satisfactory localization results compared with R-C3D in a more efficient anchor-free manner, the saliency-based module can further improve the predictions.

Besides, to visualize the channel-wise mean of the start-sensitive and end-sensitive features with and without Boundary Consistency Learning (BCL) better, we also provide a feature visualization demo video in our supplementary material.

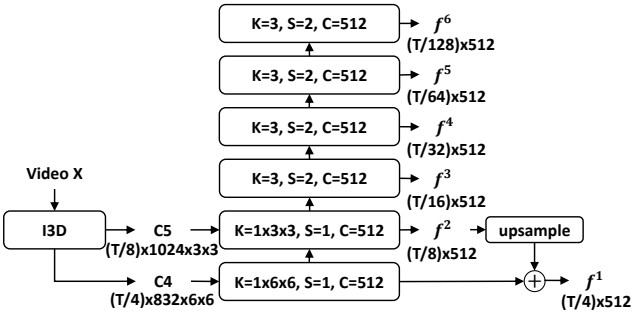


Figure 1. Detailed Architecture of our feature extractor. K, S and C denote kernel size, stride and channel of convolution layers respectively. $\oplus$ denotes element-wise summation.

Dynamic FPS	ActivityNet1.3			
	0.5	0.75	0.95	Avg.
×	50.9	33.1	6.2	32.7
✓	**52.4**	**35.3**	**6.5**	**34.4**

Table 1. Comparison of model trained with and without dynamic FPS on ActivityNet1.3.

Post Process	THUMOS14					
	0.3	0.4	0.5	0.6	0.7	Avg.
NMS	65.1	60.4	53.6	42.0	29.1	50.4
Soft-NMS	**67.3**	**62.4**	**55.5**	**43.7**	**31.1**	**52.0**

Table 2. Comparison of models trained with and without soft-NMS on THUMOS14.

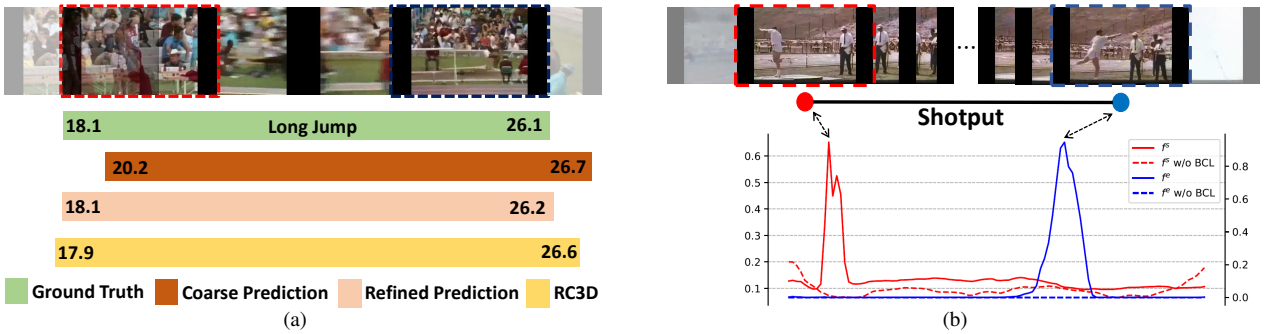
Model	THUMOS14			ActivityNet1.3		
	RGB	Flow	Fusion	RGB	Flow	Fusion
Baseline w/o GN	37.4	38.9	46.2	30.9	30.8	31.9
Baseline	40.4	40.4	48.5	31.0	31.5	32.4
+quality	41.4	41.2	49.9	31.7	32.1	33.3
+quality+refine	42.0	42.4	50.4	32.5	32.7	33.7
Full model	**43.2**	**44.2**	**52.0**	**32.9**	**33.1**	**34.4**

Table 3. Ablation results on ActivityNet1.3 and THUMOS14. Average mAP is reported.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[2] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Pro-*



(a) (b)

Figure 2. Visualization results of our model: (a) Localization example in THUMOS14. We compare the predictions among our coarse and refined predictions, RC3D result and the ground truth. All boundaries are shown in seconds. (b) The visualization of the channel-wise mean of the start-sensitive and end-sensitive features f^s, f^e and those features without BCL.

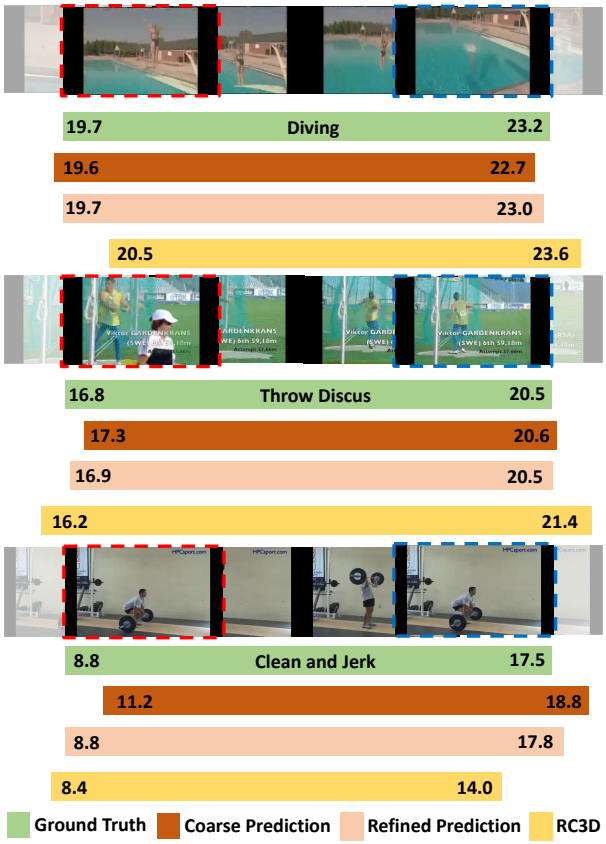


Figure 3. Visualization results among our coarse and refined prediction and anchor-based method R-C3D [2].

ceedings of the IEEE international conference on computer vision, pages 5783–5792, 2017. 1, 2, 3