Appendix A: Temporal Proposal Network

Following R-C3D [2], the temporal proposal network (TPN) in our method is an anchor based model for temporal action span prediction. The detailed structure is depicted in Table 1, from conv1 to conv4, we add two additional 1×1 convolution blocks to predict the probability of the existence of actions and regress the starting and ending timestamps (t_s, t_e) w.r.t the 1D anchor associated with the temporal position.

block	layers	input size	output size
I3D	-	$3 \times 96 \times 256 \times 256$	$832 \times 12 \times 16 \times 16$
conv1	3D conv, stride 1, kernel 1	$832 \times 12 \times 16 \times 16$	$512 \times 12 \times 16 \times 16$
	relu	$512 \times 12 \times 16 \times 16$	$512 \times 12 \times 16 \times 16$
	spatial pooling	$512 \times 12 \times 16 \times 16$	512×12
	1D conv, stride 1, kernel 1	512×12	512×12
conv2	1D conv, stride 1, kernel 1	512×12	256×12
	relu	256×12	256×12
	1D conv, stride 2, kernel 3	256×12	512×6
	relu	512×6	512×6
conv3	1D conv, stride 1, kernel 1	512×6	256×6
	relu	256×6	256×6
	1D conv, stride 2, kernel 3	256×6	512×3
	relu	512×3	512×3
conv4	1D conv, stride 1, kernel 1	512×3	256×3
	relu	256×3	256×3
	1D conv, stride 2, kernel 3	256×3	512×1
	relu	512×1	512×1

Table 1. architecture of models including details of TPN, the size is in order of $C \times T \times H \times W$

Appendix B: Analysis on Segment Matching

We analysis the sensitivity of our model w.r.t the different positive matching configuration, we reports video-mAP@0.5 on UCF101-24 dataset [1] in table 2. We test three different positions: the first K frames, the middle K frames and the last K frames. The results show that under the same segment length K, the effect of segment position is marginal. On the other hand, the performance is less sensitive when $K \leq 10$, but degrades when K is set to 14. This can be attributed to small IOU values between anchor and segment under such settings, hence the total number of positive samples is small.

segmetn length K	2	6	10	14
from beginning	62.5	62.7	62.0	59.8
from middle	62.6	62.7	62.4	60.3
from end	62.1	62.4	62.1	59.9

Table 2. Comparison between different positive matching configuration on UCF101-24

References

1. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012)
2. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: ICCV. pp. 5783–5792 (2017)