Supplementary for Paper: Learning Pixel-Level Distinctions for Video Highlight Detection

Fanyue Wei[1*] Biao Wang[2] Tiezheng Ge[2] Yuning Jiang[2] Wen Li[1] Lixin Duan[1†]

[1]School of Computer Science and Engineering & Shenzhen Institute for Advanced Study, UESTC [2]Alibaba Group

{wfanyue, liwenbnu, lxduan}@gmail.com, {eric.wb, tiezheng.gtz, mengzhu.jyn}@alibaba-inc.com

In this supplementary material, we provide the implement details, more visualization showcase examples and the detail ablation results.

1. Implement Details

We take an encoder-decoder structure for implementation. The encoder net can be constituted by any 3D Convolution Network (such as C3D [6], S3D [8] and *etc*.), while the component of decoder net differs from different encoder net in order to obtain the target highlight frame as same size as the input frame. We choose the input clip consisting by continuous frames at random to enjoy a global distribution.

In our experiments, we implement the encoder-decoder structure with backbone TASED-Net [3]. The size of input and target frame is 384×224. At training stage, it takes TASED-Net pretrained on DHF1K [7] as the saliency mask generator to be the auxiliary spatial module, the encoder net and decoder net are initialized with parameters pretrained on Kinetics [1]. While in inference stage, it only need the encoder and decoder net. We resize the input frame by 384×224 and set the number of input frames as 32 to compose an input video clip. Finally, it takes approximately six hours for the TASED-Net implementation.

For the model we takes in ablation study *Ours(C3D)* in *YouTube Highlight dataset*, we take the C3D [6] as the encoder net, while the decoder net contains four ConvTransposed Layers aiming to generate the same size 112×112 as the input target frame. For *Ours(C3D)*, the number of frames as input clip is 16 and we resize the frames as 112×112, and it takes three hours for training for *Ours(C3D)*.

For the hyper parameters in our model (*i.e*. we set β to be the threshold of the highlight map to generate the pseudo label). In *YouTube Highlight dataset* [5], *TvSum* [4] and *CoSum* [2], we simply set the value of β as 0.0005, especially for certain domain(dog, gymnastics, skiing) in *YouTube Highlight dataset* as 0.001.

We implement our model in PyTorch and train all datasets for 1500 epochs using 2 Tesla V100 cards, using SGD as the optimizer.

2. More Visualization Showcases

We also show more cases that images sampled from raw videos in YouTube Highlight [5], TvSum [4] and CoSum [2] to show the detection results of highlight. As show in Figure S1, taking skiing as an example, the video content has experienced three stages of boring, foreshadowing and climax in time sequence. In the beginning, the video content is basically a single snow scene, and the appeal of this content is very low. Then the skier sets off from the top of the slope, the skiing process begins. The main content of the video changes from a static environment to skiing, which is more attractive to people in comparison. However, since the early stage of the skiing process is mainly the pure acceleration process of the skier, the attractiveness is still not high enough. At the end, he makes some professional and difficult actions. At this time, the video content becomes more exciting than the previous acceleration stage and get the highest score.

In addition, we also present the explainable visualisation figures that represents the worth watching contents in the given video frame spatially.

3. Details about ablation studies

In this section, we provide detailed information about ablation studies on dataset *TvSum* [4] and *CoSum* [2]. We supplement the details about the ablation studies on TvSum [4] and CoSum [2].

We remove the spatial module by take the Eq. 1 to generate the pixel-level pseudo label without the auxiliary spatial module.

$$d_t(i,j) = \begin{cases} 1, & I_t \in s_h \\ 0, & I_t \in s_n \end{cases} \tag{1}$$

Moreover, We remove effect of the temporal context by duplicating the target frame I_t to fill the video clip C_t. It

*Work done during an internship at Alibaba Group

†The corresponding author

Table S1. Detailed ablation results (top-5 mAP score) on the TvSum dataset.

Topic	Ours w/o temporal	Ours w/o spatial	Ours full
BK	0.745	0.767	0.845
BT	0.764	0.778	0.809
DS	0.753	0.768	0.703
FM	0.632	0.685	0.725
GA	0.742	0.719	0.764
MS	0.822	0.839	0.872
PK	0.665	0.673	0.719
PR	0.721	0.755	0.740
VT	0.672	0.669	0.744
VU	0.773	0.756	0.791
Average	0.729	0.741	0.771

Table S2. Detailed ablation results (top-5 mAP score) on the CoSum dataset.

Topic	Ours w/o temporal	Ours w/o spatial	Ours full
BJ	0.827	0.840	0.900
BP	0.803	0.960	0.970
ET	0.583	0.738	0.817
ERC	0.958	0.950	1.000
KP	0.960	0.990	1.000
MLB	0.973	0.983	1.000
NDC	0.963	0.958	0.958
NFL	1.000	0.950	0.970
SF	0.838	0.883	1.000
SL	0.979	0.899	0.844
Average	0.889	0.915	0.946

can be deemed using only I_t for distinction estimation as described in Eq. 2.

$$\hat{d}_t(i,j) = \begin{cases} 0, & M_t(i,j) \leq \beta \\ 1, & M_t(i,j) > \beta \end{cases} \tag{2}$$

where β is a threshold.

The results in Table S1 and S2 validate both the temporal and spatial module of our methods.

We also give two cases that validate the explanation of our approach. Figure S2 (a) and (b) show the results for a dog show video and a gymnastics video, respectively. In each sub figure, the first line presents the frames sampled from the input videos, the second line shows the distinction map estimated by *TASED w/o spatial*, and the third line gives the distinction map estimated by *Ours(TASED)* (*i.e.*, *TASED full*). We observe that our model (*i.e.*, *Ours(TASED)*) is able to effectively capture the most attractive contents in the given videos. For example, the most distinctive regions (shown as the bright regions in the distinction map) are respectively the dog and the actor in two videos. Compared to *TASED w/o spatial*, the full model exhibits less noise in the background region, confirming the effectiveness of using visual saliency to guide the learning of pixel-level distinctions.

Table S3. The sensibility of the threshold β on YouTube Highlight

β	0	0.0005	0.001
YouTube	0.702	0.725	0.722

4. Effect of the saliency mask:

The objective of saliency mask is to eliminate the background noise and capture the spatial highlight. If the saliency mask neglects highlights but focuses on the background, it will degrade the performance. Table S3 shows the sensibility of β. The reason that the performance drops a bit is due to that our method with a larger β would predict higher values for background pixels, which may overlook some highlight content.

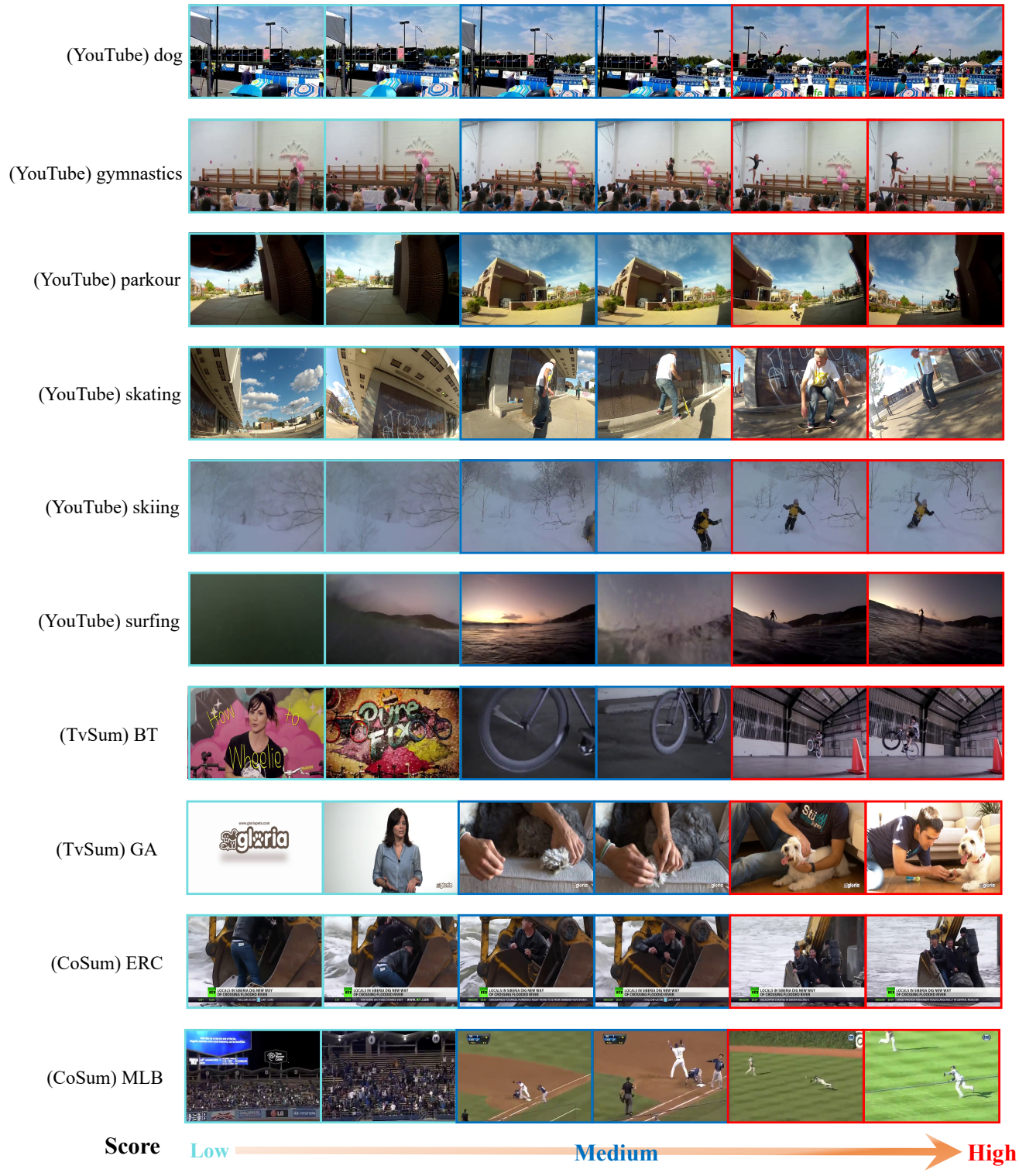

Figure S1. Showcase examples from different domain. Red means a higher highlight score, light green indicates a lower highlight score, and blue represents a medium highlight score.

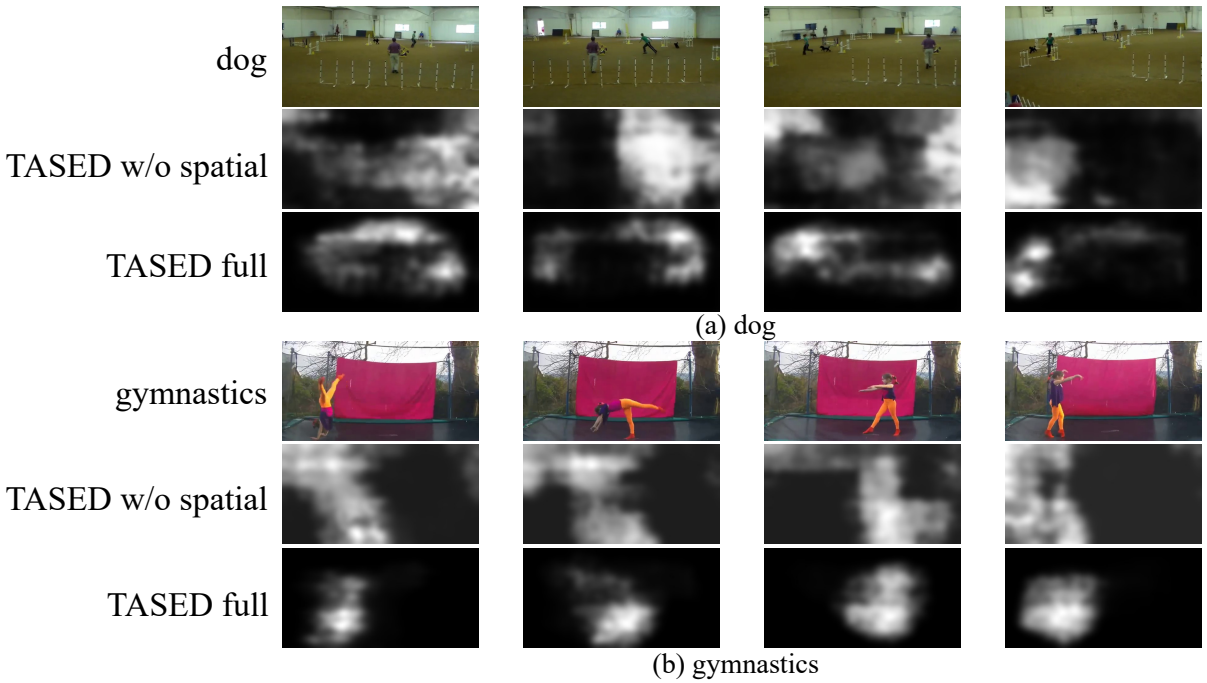

Figure S2. The frames on the first line of each subfigure are sampled from dog(gymnastics) in YouTube Highlight dataset. In the third line, the white regions inferred by our full model present the trajectory of the dog show and the action of the actress in the gymnastics clip, while the second line inferenced without the spatial module may contain some background noise and cannot provide a clear highlight cues.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1

[2] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, pages 3584–3592, 2015. 1

[3] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *ICCV*, pages 2394–2403, 2019. 1

[4] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, pages 5179–5187, 2015. 1

[5] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, pages 787–802, 2014. 1

[6] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 1

[7] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *CVPR*, pages 4894–4903, 2018. 1

[8] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018. 1