Supplementary Material for: TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos

Soufiane Belharbi[1], Ismail Ben Ayed[1], Luke McCaffrey[2], and Eric Granger[1]
[1] LIVIA, Dept. of Systems Engineering, École de technologie supérieure, Montreal, Canada
[2] Goodman Cancer Research Centre, Dept. of Oncology, McGill University, Montreal, Canada
soufiane.belharbi.1@ens.etsmtl.ca, {ismail.benayed, eric.granger}@etsmtl.ca,
luke.mccaffrey@mcgill.ca

1. CRF loss

Given an input image $\boldsymbol{X}_t$ and the softmax activation $\boldsymbol{S}_t$ of the decoder, the CRF loss is formulated [10] as,

$$\mathcal{R}(\boldsymbol{S}_t, \boldsymbol{X}_t) = \sum_{r=0}^{r=1} \boldsymbol{S}_t^{r\top} \, \boldsymbol{W} \, (\mathbf{1} - \boldsymbol{S}_t^r) \, , \qquad (1)$$

where $\boldsymbol{W}$ is an affinity matrix where $\boldsymbol{W}[i, j]$ captures the color similarity and proximity between pixels i, j in the image $\boldsymbol{X}_t$. We consider using Gaussian kernel to capture color and spatial similarities [6]. We use the permutohedral lattice [1] for fast computation of $\boldsymbol{W}$. Minimizing Eq.1 pushes the decoder to produce consistent activations for nearby pixels with similar color.

2. Classification performance

Although it is not commonly provided, we present classification performance in Tab.1 for our trained CAM-methods since they are able to do both tasks: classification, and localization. These methods yielded descent classification performance. However, there is a large margin between both datasets showing the difficulty of `YTOv2.2` dataset.

3. Visual results and demonstrative videos

Fig.1, 2 present more prediction cases over labeled ground truth frames. More illustrative videos can be downloaded from this google-drive link: `https://drive.google.com/drive/folders/1SjPed6h3XaxmwrWuYv-h9dhNhwEsRH9P?usp=sharing`.

Methods	`YTOv1`	`YTOv2.2`
CAM [11] *(cvpr,2016)*	85.3	73.9
GradCAM [9] *(iccv,2017)*	85.3	71.3
GradCAM++ [2] *(wacv,2018)*	84.4	72.4
Smooth-GradCAM++ [7] *(corr,2019)*	82.6	**75.2**
XGradCAM [3] *(bmvc,2020)*	**87.3**	71.6
LayerCAM [4] *(ieee,2021)*	84.4	72.1
TCAM (ours)	84.4	72.1

Table 1: Classification accuracy (`CL`) on test set of `YTOv1` [8] and `YTOv2.2` [5] datasets.

4. Our code

We provide our implementation code using Pytorch[1] framework[2].

Acknowledgment

This research was supported in part by the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Digital Research Alliance of Canada (alliancecan.ca).

References

[1] A. Adams, J. Baek, and M. A. Davis. Fast high-dimensional filtering using the permutohedral lattice. *Comput. Graph. Forum*, 29(2):753–762, 2010.
[2] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018.
[3] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In *BMVC*, 2020.

[1] `https://pytorch.org`
[2] Code: https://github.com/sbelharbi/tcam-wsol-video.

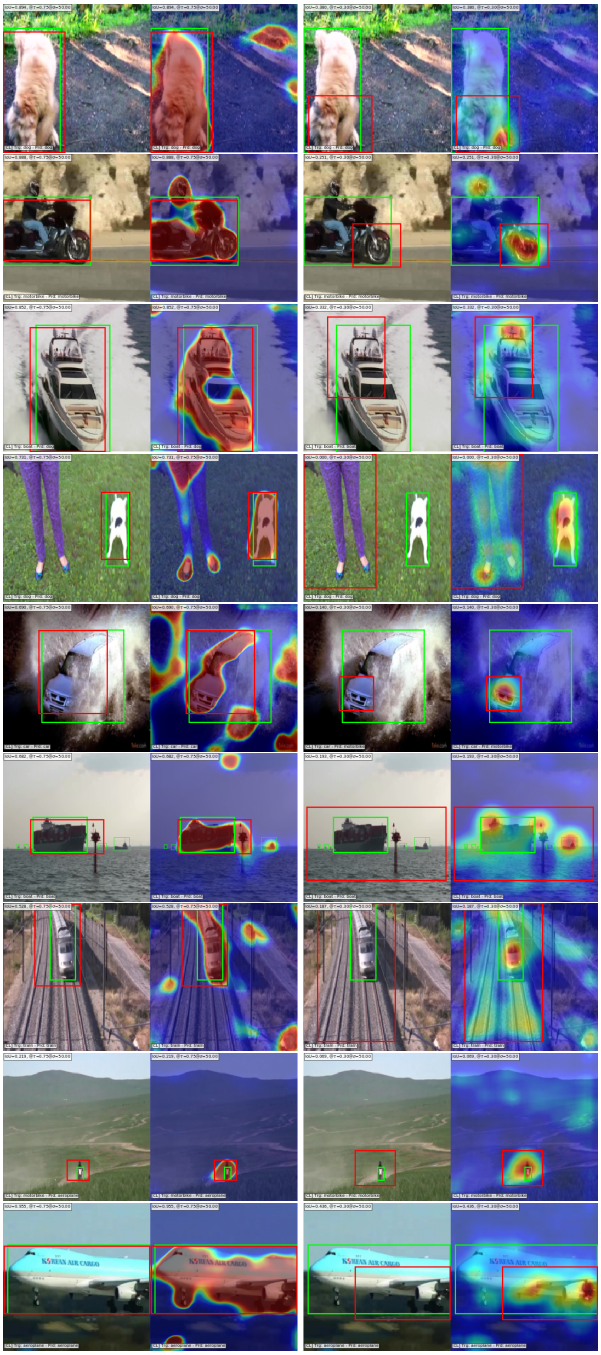

Figure 1: Prediction examples of test sets frames. *Left*: TCAM (ours). *Right*: baseline CAM method, LayerCAM [4]. *Bounding box*: ground truth (green), prediction (red). Second column is predicted CAM over image.

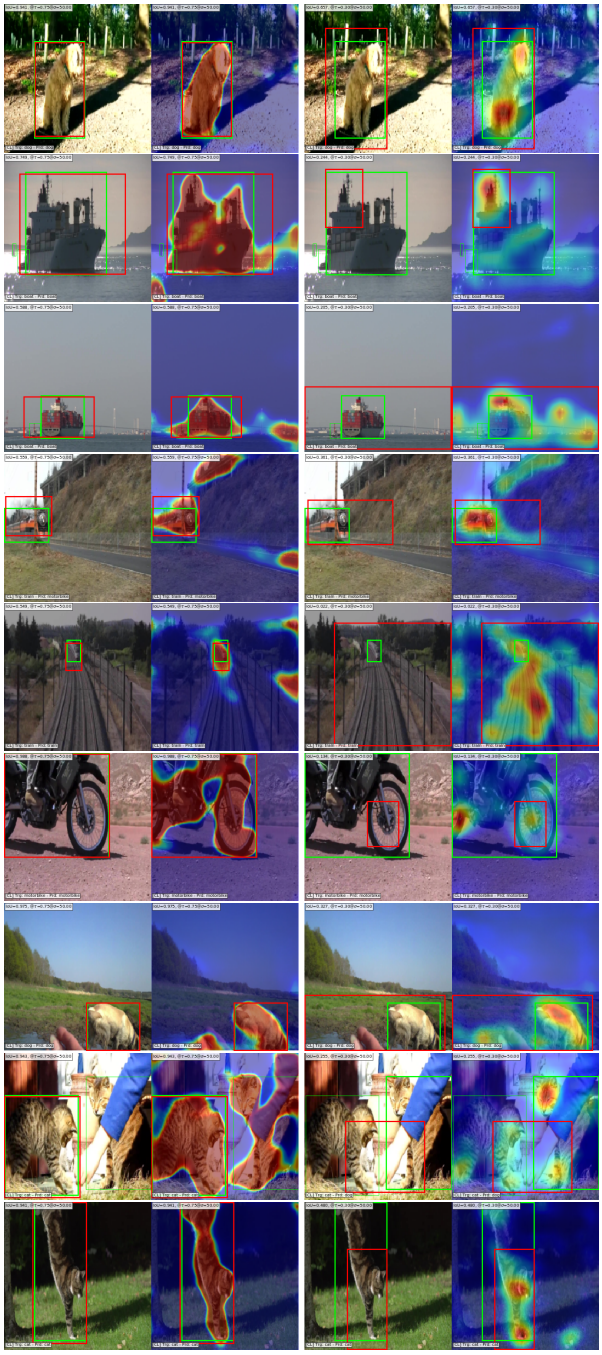

Figure 2: Prediction examples of test sets frames. *Left*: TCAM (ours). *Right*: baseline CAM method, LayerCAM [4]. *Bounding box*: ground truth (green), prediction (red). Second column is predicted CAM over image.

[4] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.*, 30:5875–5888, 2021.

[5] Vi Kalogeiton, V. Ferrari, and C. Schmid. Analysing domain shift factors between videos and images for object detection. *TPAMI*, 38(11):2327–2334, 2016.

[6] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J.S.-Taylor, R.S. Zemel, P.L. Bartlett, F.C. N. Pereira, and K.Q. Weinberger, editors, *NeurIPS*, 2011.

[7] D. Omeiza, S. Speakman, C. Cintas, and K. Weldemariam. Smooth grad-cam++: An enhanced inference level visualization technique for

deep convolutional neural network models. *CoRR*, abs/1908.01224, 2019.

[8] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[10] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov. On regularized losses for weakly-supervised cnn segmentation. In *ECCV*, 2018.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.