Supplementary Material
Discriminative Sampling of Proposals in Self-Supervised Transformers for Weakly Supervised Object Localization

Shakeeb Murtaza[1], Soufiane Belharbi[1], Marco Pedersoli[1], Aydin Sarraf[2], and Eric Granger[1]
[1] LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
[2] Ericsson, Global AI Accelerator, Montreal, Canada
{shakeeb.murtaza.1, soufiane.belharbi.1}@ens.etsmtl.ca, aydin.sarraf@ericsson.com
{marco.pedersoli, eric.granger}@etsmtl.ca

This document contains the following material:

A) an evaluation measures and error dissection;

B) an overview of CRF loss;

C) additional error analyses;

D) ablation studies;

E) additional visual results;

F) details on the TS-CAM method.

A. Performance Measures and Error Dissection

A.1. Evaluation Measure for Error Dissection

In this section, we present the evaluation measures that are used in [4] for error dissection over wrong predictions. These measures are useful for deciding threshold values for producing bounding boxes from localization maps. Specifically, localization part error (LPE) and localization more error (LME) help in deciding whether to increase or decrease the threshold value for optimal results. More details on error measures are given below:

Localization Part Error (LPE): This measure identifies that an object partially detected by the localization map with a large margin has an intersection over the predicted bounding box (IoP) > 0.5.

Localization More Error (LME): It indicates that the predicted bounding box is larger than the actual box and covers other objects or background. This can be identified if intersection over annotated-bounding-box (IoA) > 0.7.

A.2. Additional Performance Measures

Top-1 Localization Accuracy (Top-1 Loc): A prediction is considered true if the predicted class is the same as the ground truth and the intersection over Union (IoU) is greater than 0.5.

Top-5 Localization Accuracy (Top-5 Loc): A prediction is considered true if the IoU is greater than 0.5 and the actual class matches at least one of the top 5 predicted classes.

B. Overview of CRF loss

Conditional random fields (CRF) loss, aligns the predicted localization map M with the boundaries of a concerned object presented in input x. CRF loss [10] between x and M can be defined as follows:

$$\mathcal{L}_{CRF}(\mathrm{A}, \mathrm{M}) = \sum_{i=0}^{i=1} {M_i}^T A(1 - M_i) \tag{S1}$$

where A represents an affinity measure that contains mutual similarities between pixels, including proximity and color information. For capturing affinities of pixels, we use a Gaussian kernel [8] and employ permutohedral lattice [1] to reduce the computation overhead.

C. Extended Error Analysis

Further error analysis (according to the error measures defined in Section A.1) on the CUB datasets is presented in Table S1. Our method localized the correct region of the concerned object instead of overestimating or underestimating the region. It also shows that the maps generated by DiPS are very robust and have much fewer errors com-

	LPE ↓	LME ↓
VGG16	21.91	10.53
InceptionV3	23.09	5.52
TS-CAM [4]	6.30	2.85
DiPS (our)	0.05	0.07

Table S1. Extended error analysis on the CUB-200-2011 dataset

.

pared to the baseline methods. The statistics of the baseline methods are from [4].

D. Additional Ablation Studies

The performance of DiPS with various loss function combinations is shown in Table S2. It shows that all of the auxiliary losses contribute significantly towards the final performance. Also, training through arbitrary selection of pixels (pseudo-labels) rather than the classifier loss or fixed pseudo-labels allows DL models to explore different regions of an object and can provide accurate localization. Adding CRF and classification terms at the same time significantly improves the performance of our model. The `MaxBoxAcc` of our model on TelDrone is presented in table S3. Furthermore, the `MaxBoxAcc`, `top-1` and `top-5` localization accuracy for CUB dataset is presented in Table S4. We achieved state-of-the-art performance on the TelDrone and CUB dataset.

	CUB (`MaxBoxAcc`)	TelDrone (`MaxBoxAcc`)
$\mathcal{L}_{CPA} + \mathcal{L}_{CRF}$	95.4	93.3
$\mathcal{L}_{CPA} + \mathcal{L}_{CLS}$	94.6	91.7
$\mathcal{L}_{CPA} + \mathcal{L}_{CRF} + \mathcal{L}_{CLS}$	97.0	96.2

Table S2. Localization performance of our DiPS method with different losses.

	`MaxBoxAcc`
CAM [15] *(cvpr,2016)*	50.9
HaS [9] *(iccv,2017)*	60.4
ACoL [13] *(cvpr,2018)*	62.3
SPG [14] *(eccv,2018)*	67.9
ADL [3] *(cvpr,2019)*	73.5
CutMix [12] *(eccv,2019)*	54.7
DiPS (ours)	962

Table S3. `MaxBoxAcc` performance on the TelDrone dataset.

	CUB		
	`MaxBoxAcc`	`top-1` Loc Acc	`top-5` Loc Acc
CAM [15] *(cvpr,2016)*	73.2	56.1	–
HaS [9] *(iccv,2017)*	78.1	60.7	–
ACoL [13] *(cvpr,2018)*	72.7	57.8	–
SPG [14] *(eccv,2018)*	63.7	51.5	–
ADL [3] *(cvpr,2019)*	75.7	41.1	–
CutMix [12] *(eccv,2019)*	71.9	54.5	–
ICL [6] *(accv,2020)*	57.5	–	–
TS-CAM [4] *(iccv,2021)*	87.7	71.3	83.8
BR-CAM [16] *(eccv,2022)*	–	–	–
CREAM [11] *(cvpr,2022)*	90.9	71.7	86.3
BGC [7] *(cvpr,2022)*	93.1	70.8	88.0
F-CAM [2] *(wacv,2022)*	92.4	59.3	82.7
DiPS (ours)	97.0	78.8	91.3
DiPS (ours) (w/ TransFG classifier [5])	97.0	88.2	95.6

Table S4. `MaxBoxAcc`, `top-1` and `top-5` performance on the CUB dataset.

E. Visual Results

Visual representation of our method compared to baseline methods on CUB is illustrated in Fig.S2. Our method generates a very smooth map instead of hot-spotting different parts of the concerned object. Ultimately, the map generated by our method does not require an extensive threshold search to find an optimal bounding box. Compared to the `class` tokens of SST (used to harvest pseudo-labels), our method is able to learn an effective localization map from noisy pseudo-labels.

F. Details on Baseline Method: TS-CAM

By taking advantage of the attention mechanism, TS-CAM [4] is capable of capturing long-range dependency among different parts of an image. As a result, it can efficiently separate background and foreground objects. In other words, it first divides the images into a set of patches for capturing long-range dependency information and records its effects in `class` token. The attention of `class` token is then fused with the semantic aware map to produce the final attention/activation map. The flow diagram of TS-CAM is depicted in Fig.S1. Lastly, a detailed visualization of the internal representation of the token TS-CAM `class` is presented in Fig.S3. It shows that the average of all maps could potentially include noise and background regions in the final prediction.

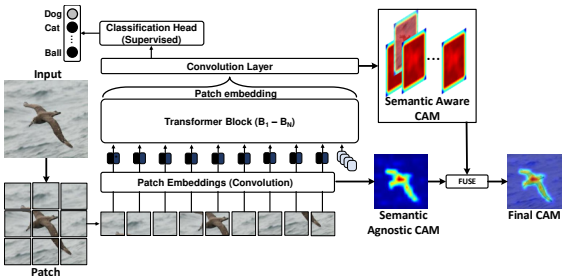

Figure S1. Illustration of the baseline Token Semantic Coupled Attention Map (TS-CAM) model for WSOL.

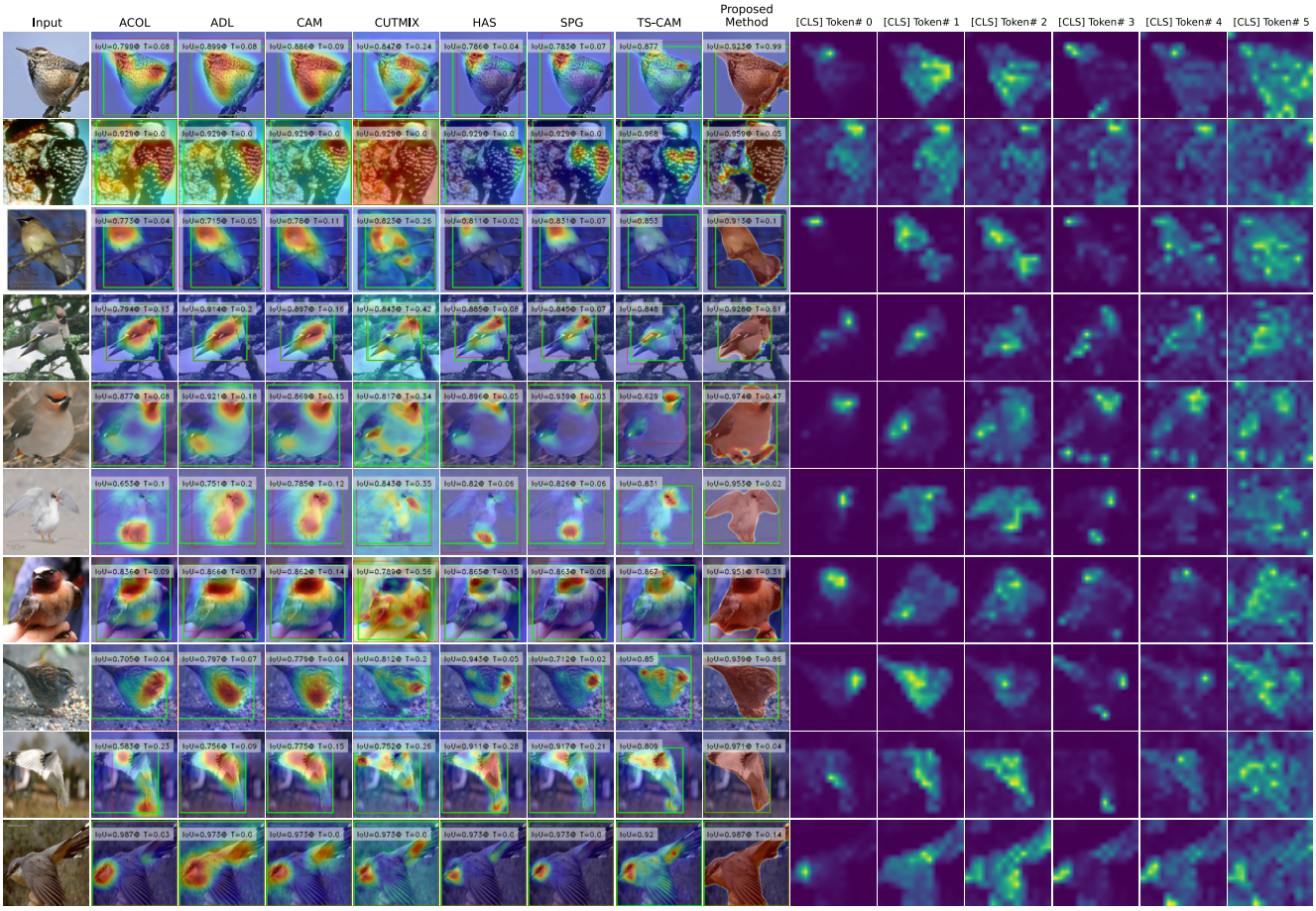

Figure S2. Examples of visual results on the CUB-200-2011 dataset.

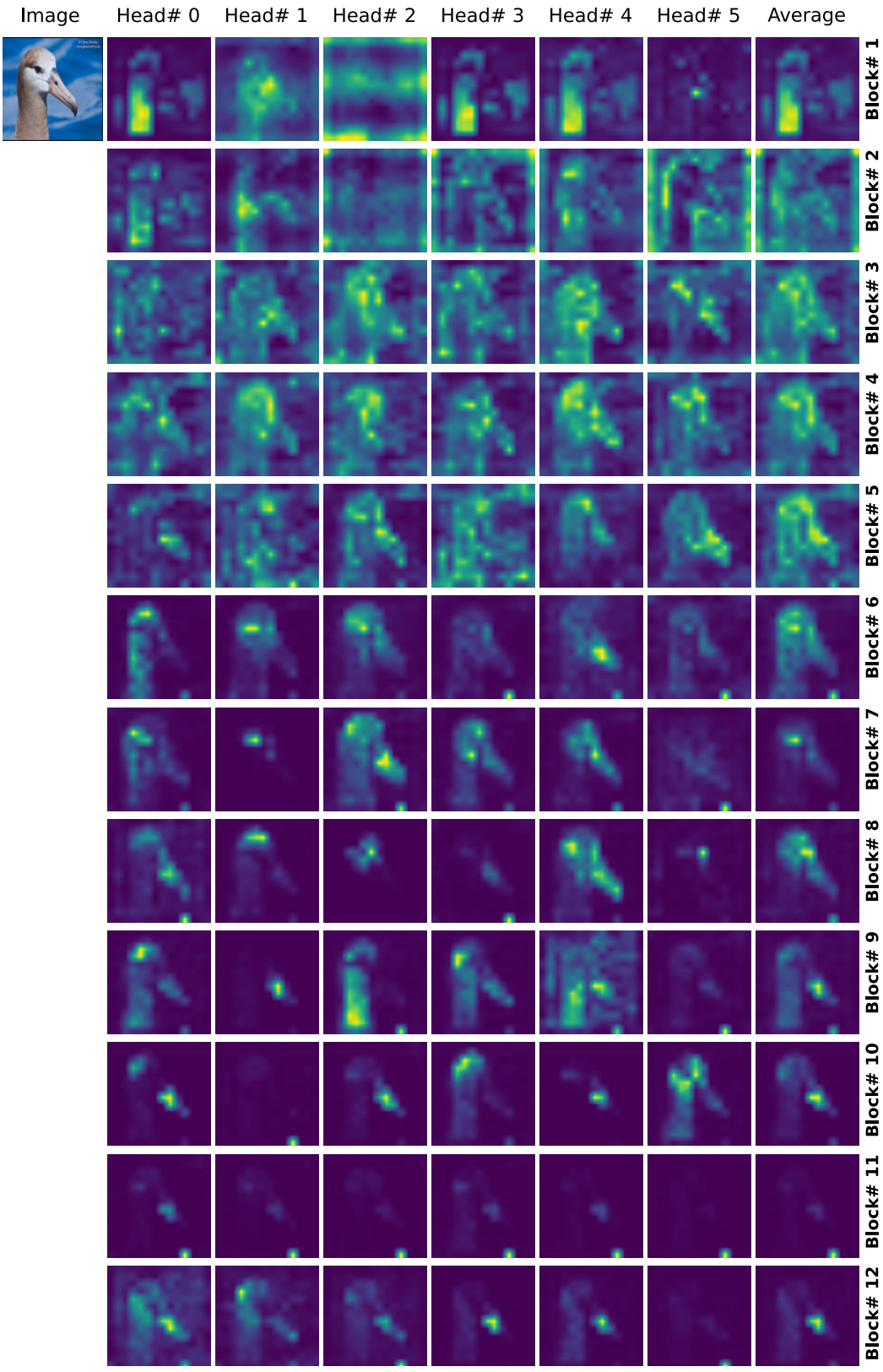

Figure S3. Attention map of each transformer head (`class` token) learned by TS-CAM. However, different parts of the object are accumulated across the layers/blocks, and it must include semantic aware CAM to suppress noise and generate final results.

References

[1] Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. In *Computer graphics forum*, volume 29, pages 753–762. Wiley Online Library, 2010.

[2] S. Belharbi, A. Sarraf, M. Pedersoli, I. Ben Ayed, L. McCaffrey, and E. Granger. F-cam: Full resolution class activation maps via guided parametric upscaling. In *WACV*, 2022.

[3] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

[4] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021.

[5] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 852–860, 2022.

[6] Minsong Ki, Youngjung Uh, Wonyoung Lee, and Hyeran Byun. In-sample contrastive learning and consistent attention for weakly supervised object localization, 2020.

[7] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14258–14267, 2022.

[8] Vladlen Koltun et al. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*, 4, 2011.

[9] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017.

[10] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018.

[11] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. Cream: Weakly supervised object localization via class re-activation mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9437–9446, 2022.

[12] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[13] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018.

[14] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 597–613, 2018.

[15] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[16] Lei Zhu, Qian Chen, Lujia Jin, Yunfei You, and Yanye Lu. Bagging regional classification activation maps for weakly supervised object localization. *arXiv preprint arXiv:2207.07818*, 2022.