Supplementary Material: A Realistic Protocol for Evaluation of Weakly Supervised Object Localization

Shakeeb Murtaza, Soufiane Belharbi, Marco Pedersoli, Eric Granger
LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada

This supplementary material contains the following content:

- **A. Review of evaluated methods and their search spaces.** In this section, we provided a detailed description of methods employed to evaluate the efficacy of our proposed protocol for model selection. Additionally, it outlines various hyperparameters for each method and their respective feasible range.
- **B. Generation of noisy ground truth bboxes.** This section delineates the methodology for generating noisy ground truth (GT) bounding boxes (bboxes) to analyze their impact on model selection.
- **C. Pseudo-bboxes performance across different selection steps.** This section presents the performance of pseudo-bboxes at different selection steps of our proposed method.
- **D. Localization evaluation in WSOL (thresholded-`IoU` vs `IoU`).** This section describes an ongoing challenge in WSOL where discrepancies in localization performance exist between the commonly used `MaxBoxAcc` and the `IoU` metric.
- **E. Localization Evaluation with different thresholds.** This section presents the results at different thresholds along with their average.
- **F. Experiments with medical datasets.** To show the generalizability of our proposed evaluation protocol across various domains, we employ medical datasets.

A. Review of Evaluated Methods and their Search Spaces

A.1. Evaluated Methods

To assess the efficacy of the proposed protocol, we employ eight methods published in top-tier venues from 2016 to 2023. These methods are presented chronologically within this section, providing a comprehensive overview.

Class activation mapping (CAM) [12] is able to extract an activation map for a particular class using a pre-trained CNN-classifier with global average pooling. It generates the final map by aggregating different activation maps from the penultimate convolution layer based on the contribution towards each class by using weights from the last fully connected layer.

Hide-and-seek (HaS) [7] force the network to look beyond discriminative regions of a particular object by augmenting the input image. It hides patches of input image during training by employing two hyper-parameters; drop rate and grid size.

Adversarial complementary learning (ACoL) [11] employees an architecture with two parallel classifier heads that tires to find complementary regions by adversarially erasing high-scoring activations.

Attention-based dropout layer (ADL) [4] works similarly as ACoL by erasing high-score activation to force by employing drop masks generated without second classifiers head.

Non-local combinational class activation maps (NL-CCAM) [9]. In this paper, the author argues that employing the activation map of the class with the highest classifier's score may only highlight discriminative regions and for certain images it tends to focus on background regions. To address this limitation, the author proposes to combine activation maps of different classes. This combination is based on the respective probability score of each, encompassing a spectrum from the highest to the lowest.

Token semantic coupled attention map (TS-CAM) [5]. TS-CAM is a cascaded ViT-CNN architecture that proposes to redistribute class information to patch tokens. This is achieved by implementing a CNN-based classification (`CL`) head atop the patch tokens, thereby rendering the `CLS` token class-agnostic. Therefore, these `CLS` tokens are combined with the activation map extracted from the last convolutional layer to produce an activation map highlighting different object parts of a particular class.

Spatial calibration module (SCM) [1]. This paper introduced an SCM module atop the transformer features to align the boundaries of the generated map with the object boundaries by avoiding partial activation in different areas of the activation map. This module integrates semantic similarities presented in patch tokens and their spatial relationships into a unified model. SCM effectively recalibrates the trans-

former's attention and semantic representations to mitigate the background noise and sharpen object boundaries.

Task-specific spatial-aware token (SAT). [8] This paper introduces a spatial-aware token (SAT) into the transformer's input space. Like `CLS` token that is able to accumulate information for `CL` tasks, it is incorporated to aggregate the global representation of the object of interest. Furthermore, the SAT is a passed-to spatial-query attention module that treats the SAT as a query to calculate similarity with different patches and produces probabilities for foreground object for producing accurate localization maps.

A.2. Hyperparamter Search Space

To fairly compare different WSOL methods, we took steps to minimize human biases during training. This includes employing pseudo-bboxes for the evaluation and sampling hyperparameter values from the feasible range, except for annotations on the test set, which were used only to assess the trained model's performance.

Each method was trained using four shared hyperparameters, while the additional hyperparameters were specific to each model. We sampled the values for these hyperparameters from their feasible range. A cartesian product of these and shared hyperparameters was computed to create the final grid of hyperparameters for training each model. A detailed summary of the hyperparameters employed to train different WSOL models is presented in Tab.S1.

Method	Hyperparameter	Sampling Distribution	Range
Common HPs	LR, WD, Gamma	LogUniform	$[10^{-5}, 10^{0}]$
	Step Size	Uniform	CUB: $[5-45]$ ILSVRC: $[2-9]$
CAM [12], TS-CAM [5] SCM [1], NL-CCAM [9]	Common HPs	-	-
HaS [7]	Drop Rate, Drop Area	Uniform	$[0, 1]$
ACoL [11]	Erasing Threshold	Uniform	$[0, 1]$
ADL [4]	Drop Rate, Erasing Threshold	Uniform	$[0, 1]$
SAT [8]	Area Threshold	Uniform	$[0, 1]$

Table S1. Hyperparameter search space for different methods

Method	CUB (IoU)			ILSVRC (IoU)		
	SS	RPN	CLIP	SS	RPN	CLIP
Mean IoU (PG*)	32.21	28.71	–	23.38	27.57	–
Mean IoU (PG*, 20% Filtered)	33.69	37.89	–	34.99	37.17	–
Mean IoU (Top Box, 20% by Objectness Score)	39.90	69.80	–	44.90	54.02	–
Mean IoU (Top Box, 20%, PG*+Scoring)	39.98	71.23	–	45.07	61.08	–
Upper Bound (Select by IoU with GT)	64.07	83.66	–	65.46	84.42	–
IoU from Otsu	–	–	68.80	–	–	64.41
IoU using 1K Threshold	–	–	69.56	–	–	65.78

*PG: Pointing Game

Table S2. Performance of pseudo-bboxes obtained using different off-the-shelf region proposal methods across multiple refinement stages over the validation set. The table highlights the incremental improvement in `IoU` through various selection steps.

B. Generation of Noisy Ground Truth Bboxes

In the main paper, we introduce a validation protocol designed to evaluate the robustness of the proposed model selection techniques in the presence of noisy GT bboxes. This protocol systematically perturbs the GT bboxes, initially used for model selection, to emulate conditions of noisy or imprecise GT annotations. To produce noisy GT bboxes, a sequence of random transformations is applied to the GT bboxes, creating varying noise levels. For each transformation, a total of ten unique noise levels are defined. These levels signify the maximum likelihood of deformation at each noise level. This likelihood is derived by sampling the deformation value using a uniform distribution that varies from 5 to 50, with intervals set at 5. To generate nosy bboxes we, first apply, scaling transformation to the GT bboxes between -50% and +50% with a maximum likelihood of a particular noise level. Following the scaling, we apply shift transformation to the scaled bbox by choosing a random shift length, where the shift length is set between 0% and the maximum size percentage corresponding to a particular noise level. Finally, we modify the aspect ratio of bbox based on a probability factor 'p' which indicates the likelihood, representing a specific noise level.

C. Pseudo-bboxes Performance Across Different Selection Steps

Different off-the-shelf models, SS, RPN, and CLIP, are employed to produce pseudo-bboxes. These methods generate a set of class-agnostic pseudo-bboxes. To select discriminative boxes from a pool of object proposals, the pointing game analysis was employed [10]. This involves harvesting CAM from a pre-trained classifier and pinpointing the peak activation that is used to select discriminative boxes. Despite the initial filtering of bboxes via the pointing game, a substantial number of bboxes remained. To address this, we employ a sequential refinement process, in which we initially filter the top 20% based on objectness or classifier score for boxes obtained from RPN and SS, respectively. Subsequently, the pointing game was employed to refine this selection, followed by selection of top-performing boxes based on score. In the case of CLIP, we utilized Otsu's thresholding method to identify binary maps, upon which bboxes were delineated around the largest connected areas. A comprehensive description of the proposed method for generating pseudo-bboxes is provided in the main paper.

The performance of pseudo-bboxes at different selection steps is presented in Tab.S2. This table shows that as we select relevant bboxes generated by SS or RPN at each stage, we progressively choose better-performing bboxes, resulting in reliable performance relative to the upper bound performance when using GT bboxes to select top-performing bbox. Initially, pseudo-bboxes generated by SS and RPN

Method	Backbone	CUB-200-2011 (MaxBoxAcc)					CUB-200-2011 (IoU)					ILSVRC (MaxBoxAcc)					ILSVRC (IoU)				
		CL	GT	RPN	CLIP	SS	CL	GT	RPN	CLIP	SS	CL	GT	RPN	CLIP	SS	CL	GT	RPN	CLIP	SS
CAM [12] *(cvpr,2016)*	ResNet50	66.98	70.40	71.10	70.62	69.89	55.53	56.71	56.88	56.65	56.76	61.48	64.06	63.60	63.90	63.60	56.17	57.89	57.42	57.68	57.42
HaS [7] *(iccv,2017)*	ResNet50	67.62	75.85	75.85	75.85	74.73	57.01	59.81	59.81	59.81	59.39	61.69	63.77	63.94	63.77	63.30	56.81	58.49	58.39	58.49	57.32
ACoL [11] *(cvpr,2018)*	ResNet50	66.62	74.64	74.64	75.37	74.14	55.32	58.29	58.29	58.55	58.13	61.98	62.93	62.75	63.45	62.92	55.62	56.39	56.32	56.94	56.39
ADL [4] *(cvpr,2019)*	ResNet50	67.82	76.63	76.63	76.06	74.99	55.64	59.12	59.12	58.93	58.32	62.81	65.11	65.97	65.11	65.19	56.39	58.46	58.37	58.55	58.54
NL-CCAM [9] *(wacv,2020)*	VGG-GAP	64.15	65.58	65.58	65.22	45.97	54.11	54.76	54.76	54.59	47.88	58.42	60.63	60.63	60.63	52.72	51.59	54.98	54.98	54.98	49.44
TS-CAM [5] *(iccv,2021)*	DeiT-S	88.36	90.19	90.35	89.52	88.71	69.14	69.78	69.83	69.95	68.36	56.40	66.75	66.75	66.75	66.17	53.67	59.54	59.54	59.54	59.00
SCM [1] *(eccv,2022)*	DeiT-S	88.47	91.56	92.25	92.26	91.76	68.64	70.27	70.89	70.93	70.34	57.92	61.76	61.75	61.75	59.76	52.13	54.55	54.56	54.56	53.47
F-CAM [2] *(wacv,2022)*	ResNet50	24.95	89.83	89.23	89.81	88.26	37.72	68.79	68.12	68.62	68.30	–	–	–	–	–	–	–	–	–	–
SAT [8] *(iccv,2023)*	DeiT-S	79.70	92.14	92.45	91.45	92.23	63.13	73.67	73.59	72.92	73.61	64.94	70.12	67.08	70.13	70.13	56.09	62.80	58.55	62.80	62.80

Table S3. Comparative Analysis of MaxBoxAcc (IoU-50) versus IoU on CUB and ILSVRC with different model selection methods.

are filtered based on objectness or classifier scores, followed by a pointing game analysis for further refinement. The results indicate a significant improvement in the mean intersection over union (IoU) across these selection stages. For example, after the initial selection and filtering of the top 20% based on objectness scores, the IoU increases substantially, demonstrating the efficacy of our proposed method. In contrast, CLIP generates activation maps that highlight particular objects. Otsu's thresholding method is employed to convert these maps into binary images, enabling the delineation of bboxes around the largest connected areas. Despite the single-stage selection process for CLIP-generated maps, the resulting bounding boxes achieve competitive performance.

D. Localization Evaluation in WSOL (Thresholded-IoU vs IoU)

So far, we have reported the localization performance using MaxBoxAcc metric [3]. It is also known as as *GT-known localization* metric. It scores one point when the IoU between the GT bbox and the predicted box is above 50%, otherwise, it scores 0. It is referred to IoU 5-0 as well. It is a well established and commonly used metric in WSOL. In addition to IoU-50, we report the IoU in Tab.3 of the main paper. These results show that model selection using CL accuracy still lead to poor IoU. However, selection using our proposed pseudo-bboxes yields competitive IoU compared to when using oracle bboxes.

The comparison based on IoU (Tab.3 of the main paper) lead us to an interesting result presented in Tab.S3. This table shows that MaxBoxAcc, using the commonly oracle bboxes, gives largely higher localization scores compared to the exact localization accuracy reported by IoU. For instance, SAT method [8] scores 92.14% in MaxBoxAcc, while it only scores 73.12% over CUB dataset. When considering only MaxBoxAcc, the results give the impression that CUB dataset is saturated, especially when the same authors [8] have reported a MaxBoxAcc of 98.45%. However, when inspecting IoU metric, localization is still low at 73.12%.

In addition, since MaxBoxAcc is based on thresholding, extreme localization scores can hit the same scoring point. For instance, a prediction with $\texttt{IoU} = 50.1\%$ scores the same point as when the prediction is $\texttt{IoU} = 99.99\%$. However, both $\texttt{IoU} = 49.9\%$, $\texttt{IoU} = 1\%$ scores 0 point in MaxBoxAcc. This makes localization evaluation less efficient.

Despite its common usage in the literature, the aforementioned limitations of MaxBoxAcc suggest that reporting IoU along with MaxBoxAcc could be beneficial in better assessing localization performance of different methods in WSOL.

Since CUB dataset is relatively easier than ILSVRC dataset, the latter paints a realistic evaluation of the progress that has been done in WSOL. Since the work of Zhou et al. [12] in 2016 up to now, only $\approx 6\%$, and $\approx 4.9\%$ of improvement has been done in term of MaxBoxAcc and IoU, respectively. This suggests that a lot of work is still needs to be done to furthermore improve WSOL methods.

E. Localization Evaluation with different thresholds

We have previously evaluated localization performance using the MaxBoxAcc metric, which assigns a score of one when the IoU between the GT bbox and the predicted bounding box exceeds 50%, and a score of zero otherwise, as well as using raw IoU without any threshold. In addition, we extend the evaluation of MaxBoxAcc by examining performance trends at multiple IoU thresholds, namely IoU-30, IoU-50, and IoU-70, along with their average, termed MaxBoxAccV2, a metric commonly used in the weakly supervised object localization (WSOL) literature [3]. The performance results for these thresholded IoU metrics and their average are detailed in Tab.S4&S5. In contrast to raw IoU, the thresholded IoU metrics demonstrate consistent trends when compared to MaxBoxAcc, which are sufficient to offer meaningful insights into the results.

F. Experiments with medical datasets

To show the robustness and generalizability of our evaluation protocol across datasets with varying characteristics, we extended our protocol to the task of localization in histology images—a particularly challenging problem due to its complexity and sensitivity, especially for non-experts attempting to identify regions of interest. For this,

		CL				GT				RPN				CLIP				SS			
Method	**Select**	IoU-30	IoU-50	IoU-70	MaxBoxV2	IoU-30	IoU-50	IoU-70	MaxBoxV2	IoU-30	IoU-50	IoU-70	MaxBoxV2	IoU-30	IoU-50	IoU-70	MaxBoxV2	IoU-30	IoU-50	IoU-70	MaxBoxV2
CAM [12] *(cvpr'16)* ResNet50	BT-TT	73.76	32.99	6.161	37.63	93.95	70.40	20.05	61.46	94.30	71.10	19.74	61.71	94.33	70.62	19.46	61.47	94.06	69.89	20.38	61.44
	BT-VT	–	–	–	–	93.13	69.50	19.07	60.56	94.25	70.43	19.22	61.30	93.73	69.88	18.46	60.69	70.05	27.97	12.80	36.94
	BV-TT	69.72	29.37	5.143	34.74	93.95	70.40	20.05	61.46	93.95	70.40	20.05	61.46	94.70	69.76	18.46	60.97	90.05	61.20	16.44	55.89
	BV-VT	–	–	–	–	93.61	69.20	19.33	60.71	93.61	68.93	19.03	60.52	94.70	67.98	16.58	59.75	58.66	22.79	15.84	32.43
HaS [7] *(iccv'17)* ResNet50	BT-TT	71.48	37.00	11.25	39.91	94.59	75.85	28.49	66.31	94.59	75.85	28.49	66.31	94.59	75.85	28.49	66.31	94.64	74.73	27.42	65.59
	BT-VT	–	–	–	–	92.33	74.68	29.15	65.38	94.44	75.69	28.27	66.13	94.14	75.69	27.71	65.84	67.22	25.42	12.08	34.90
	BV-TT	66.15	30.11	8.715	34.99	93.16	75.49	30.18	66.27	93.57	74.92	29.20	65.89	94.87	73.38	23.76	64.00	92.85	72.69	30.53	65.35
	BV-VT	–	–	–	–	92.33	74.68	29.15	65.38	93.49	74.02	28.99	65.5	94.51	73.24	23.45	63.73	61.28	22.43	27.09	36.93
ACoL [11] *(cvpr'18)* ResNet50	BT-TT	89.48	46.59	7.490	47.85	96.35	74.64	20.81	63.93	96.35	74.64	20.81	63.93	97.20	75.37	20.29	64.28	96.85	74.14	19.41	63.46
	BT-VT	–	–	–	–	95.73	73.50	20.65	63.29	96.25	73.83	20.59	63.55	96.87	74.83	19.60	63.76	88.29	33.43	9.164	43.62
	BV-TT	80.47	34.89	6.161	40.50	96.35	74.64	20.81	63.93	96.35	74.64	20.81	63.93	97.42	72.29	16.87	62.19	93.71	56.04	8.560	52.77
	BV-VT	–	–	–	–	95.73	73.50	20.65	63.29	96.25	73.83	20.59	63.55	97.25	72.21	16.60	62.01	88.60	27.85	6.817	41.08
ADL [4] *(cvpr'19)* ResNet50	BT-TT	81.08	40.97	7.973	43.34	96.47	76.63	23.14	65.41	96.47	76.63	23.14	65.41	95.35	76.06	23.16	64.85	95.46	74.99	21.86	64.10
	BT-VT	–	–	–	–	96.27	76.21	22.62	65.03	96.20	76.52	23.00	65.24	95.21	75.92	22.73	64.61	76.85	27.39	10.01	38.08
	BV-TT	77.89	37.17	6.144	40.40	95.40	75.83	23.16	64.79	93.63	75.97	27.25	65.61	96.63	75.80	20.31	64.24	75.45	32.08	6.040	37.85
	BV-VT	–	–	–	–	94.87	72.73	22.69	63.43	93.04	75.85	24.02	64.30	96.08	74.73	19.57	63.46	64.03	22.21	5.505	30.58
CCAM [9] *(wacv'20)* VGG	BT-TT	90.36	52.67	12.92	51.98	91.49	65.58	19.38	58.81	91.49	65.58	19.38	58.81	91.54	65.22	18.89	58.54	76.68	45.97	17.77	46.80
	BT-VT	–	–	–	–	88.74	64.44	18.55	57.24	90.81	65.27	19.12	58.4	90.42	62.54	18.55	57.17	72.02	31.89	15.15	39.68
	BV-TT	90.36	52.67	12.92	51.98	91.49	65.58	19.38	58.81	91.49	65.58	19.38	58.81	91.54	65.22	18.89	58.54	76.68	45.97	17.77	46.80
	BV-VT	–	–	–	–	90.12	62.44	18.77	57.11	91.37	64.84	19.27	58.49	90.42	62.54	18.55	57.17	72.02	31.89	15.15	39.68
TS-CAM [5] *(iccv'21)* DeiT-S	BT-TT	85.07	51.77	20.34	52.39	99.22	90.19	55.31	81.57	99.15	90.35	55.16	81.55	98.96	89.52	55.91	81.46	98.92	88.71	51.25	79.62
	BT-VT	–	–	–	–	99.17	89.33	51.89	80.13	99.03	90.16	54.21	81.13	98.60	89.52	55.10	81.07	66.13	29.49	13.87	36.49
	BV-TT	72.41	39.17	13.72	41.76	98.96	89.52	55.91	81.46	98.96	89.52	55.91	81.46	99.11	89.16	53.50	80.58	95.27	77.39	40.90	71.18
	BV-VT	–	–	–	–	98.56	88.85	53.40	80.27	98.87	88.90	51.58	79.78	98.44	88.95	50.98	79.45	59.52	24.19	22.19	35.30
SCM [1] *(eccv'22)* DeiT-S	BT-TT	62.49	30.82	8.767	34.02	99.25	91.56	56.10	82.30	99.30	92.25	58.40	83.31	99.36	92.26	58.49	83.37	99.24	91.76	56.50	82.5
	BT-VT	–	–	–	–	98.96	90.26	55.35	81.52	99.20	92.00	55.41	82.20	98.94	92.16	58.24	83.11	63.54	26.44	16.86	35.61
	BV-TT	62.49	30.82	8.767	34.02	99.25	91.56	56.10	82.30	99.30	92.25	58.40	83.31	99.36	92.26	58.49	83.37	97.39	80.16	35.86	71.13
	BV-VT	–	–	–	–	98.96	90.26	55.35	81.52	99.20	92.00	55.41	82.20	98.94	92.16	58.24	83.11	63.53	22.02	10.06	31.87
SAT [8] *(iccv'23)* DeiT-S	BT-TT	91.76	69.60	37.78	66.38	99.37	92.14	66.63	86.04	99.24	92.45	66.79	86.16	99.15	91.45	64.29	84.96	99.30	92.23	66.67	86.06
	BT-VT	–	–	–	–	99.36	91.66	67.34	86.12	99.03	92.26	62.82	84.70	98.49	91.00	64.03	84.50	70.65	41.62	28.04	46.77
	BV-TT	92.33	70.24	39.02	67.19	99.17	91.33	65.06	85.18	99.17	91.75	68.31	86.41	99.13	89.97	60.57	83.22	99.17	91.33	65.06	85.18
	BV-VT	–	–	–	–	98.87	90.86	65.06	84.93	98.79	91.71	66.44	85.64	97.82	89.93	60.40	82.71	77.59	28.04	14.68	40.10

Table S4. Test-set MaxBoxAccV2 that is average of three threshold IoU-30, IoU-50, IoU-70 (here MaxBoxAccV2 is aerage at three threhold) along with the it of WSOL models with different selection criteria on CUB. The *select* column presents (i) BT and BV indicate model selection based on hyperparameter configurations using the test set and validation set, respectively; (ii) TT and VT indicate that the threshold τ is selected using either the test set or validation set. For model selection on the validation set, we consider the GT as a reference, a selection based on the CL performance and the three different pseudo-bboxes generation proposed in this work: RPN, CLIP and SS. Our results for models selected with pseudo-bboxes are comparable to those of GT.

		CL				GT				RPN				CLIP				SS			
Method	**Select**	IoU-30	IoU-50	IoU-70	MaxBoxV2	IoU-30	IoU-50	IoU-70	MaxBoxV2	IoU-30	IoU-50	IoU-70	MaxBoxV2	IoU-30	IoU-50	IoU-70	MaxBoxV2	IoU-30	IoU-50	IoU-70	MaxBoxV2
CAM [12] *(cvpr'16)* ResNet50	BT-TT	73.16	46.29	20.86	46.77	81.99	64.06	40.22	62.09	81.59	63.60	39.55	61.58	82.02	63.90	39.98	61.96	81.59	63.60	39.55	61.58
	BT-VT	–	–	–	–	81.98	63.90	39.59	61.82	79.49	62.89	37.64	60.00	81.09	63.88	39.85	61.60	68.59	46.97	39.84	51.80
	BV-TT	73.81	46.83	21.26	47.30	81.99	64.06	40.22	62.09	81.59	63.60	39.55	61.58	82.06	64.01	40.02	62.03	81.59	63.60	39.55	61.58
	BV-VT	–	–	–	–	81.88	63.88	39.94	61.9	79.49	62.89	37.64	60.00	81.09	63.88	39.85	61.60	70.13	45.91	39.51	51.84
HaS [7] *(iccv'17)* ResNet50	BT-TT	65.09	37.82	16.15	39.68	81.3	63.77	42.08	62.38	81.54	63.94	41.7	62.39	81.3	63.77	42.08	62.38	81.43	63.30	39.26	61.33
	BT-VT	–	–	–	–	81.16	63.86	41.67	62.23	78.59	62.08	37.35	59.34	80.35	63.28	40.95	61.52	68.92	49.02	37.13	51.69
	BV-TT	72.59	45.80	21.01	46.46	81.54	63.94	41.7	62.39	81.55	63.34	39.77	61.55	81.82	63.54	39.56	61.63	81.43	63.30	39.26	61.33
	BV-VT	–	–	–	–	81.16	63.86	41.67	62.23	78.59	62.08	37.35	59.34	80.36	63.06	39.21	60.87	68.56	48.21	39.23	52.0
ACoL [11] *(cvpr'18)* ResNet50	BT-TT	77.99	50.46	22.76	50.40	81.34	62.93	38.23	60.83	81.89	62.75	37.56	60.73	82.21	63.45	39.13	61.59	82.14	62.92	37.50	60.85
	BT-VT	–	–	–	–	81.83	63.48	39.11	61.47	79.94	61.61	36.75	59.43	80.38	63.47	39.11	60.98	76.95	52.38	33.11	54.14
	BV-TT	77.99	50.46	22.76	50.40	82.23	63.70	39.03	61.65	81.89	62.75	37.56	60.73	82.14	62.93	37.5	60.85	82.14	62.92	37.50	60.85
	BV-VT	–	–	–	–	81.83	63.48	39.11	61.47	80.45	61.31	35.89	59.21	81.37	62.80	37.43	60.53	76.95	52.38	33.11	54.14
ADL [4] *(cvpr'19)* ResNet50	BT-TT	67.71	39.80	16.84	41.44	82.67	65.11	41.63	63.13	82.51	65.97	41.70	63.39	82.51	65.11	41.63	63.08	82.64	65.19	41.92	63.25
	BT-VT	–	–	–	–	82.66	65.2	41.49	63.11	80.68	64.29	37.49	60.82	81.65	65.04	41.33	62.67	69.47	49.61	41.59	53.55
	BV-TT	70.90	44.23	20.57	45.23	82.67	65.28	41.88	63.27	82.64	65.18	41.92	63.24	82.69	65.32	41.79	63.26	82.58	65.06	41.28	62.97
	BV-VT	–	–	–	–	82.66	65.2	41.49	63.11	80.41	62.30	37.63	60.11	81.65	65.04	41.33	62.67	72.96	47.09	41.32	53.79
CCAM [9] *(wacv'20)* VGG	BT-TT	73.84	49.80	24.99	49.54	77.84	60.63	38.39	58.95	77.84	60.63	38.4	58.95	77.83	60.63	38.39	58.95	72.34	52.72	31.22	52.09
	BT-VT	–	–	–	–	77.72	60.52	38.02	58.75	77.19	60.52	37.76	58.49	77.44	60.56	38.17	58.72	67.94	46.63	31.16	48.57
	BV-TT	73.84	49.80	24.99	49.54	77.84	60.63	38.39	58.95	77.84	60.63	38.4	58.95	77.83	60.63	38.39	58.95	72.34	52.72	31.22	52.09
	BV-VT	–	–	–	–	77.72	60.52	38.02	58.75	77.19	60.52	37.76	58.49	77.44	60.56	38.17	58.72	67.94	46.63	31.16	48.57
TS-CAM [5] *(iccv'21)* DeiT-S	BT-TT	12.45	6.476	2.144	7.023	82.91	66.75	43.77	64.47	82.91	66.75	43.77	64.47	82.91	66.75	43.77	64.47	82.56	66.17	42.87	63.86
	BT-VT	–	–	–	–	82.74	66.65	43.77	64.38	79.95	65.12	41.18	62.08	80.42	65.32	43.55	63.09	67.53	47.63	35.03	50.06
	BV-TT	12.45	6.476	2.144	7.023	82.91	66.75	43.77	64.47	82.52	66.14	42.94	63.86	82.91	66.75	43.77	64.47	73.60	55.7	35.03	54.77
	BV-VT	–	–	–	–	82.74	66.65	43.77	64.38	80.68	64.97	42.80	62.81	80.42	65.32	43.55	63.09	66.87	46.8	34.08	49.25
SCM [1] *(eccv'22)* DeiT-S	BT-TT	74.53	47.55	17.72	46.6	80.60	61.76	34.44	58.93	80.60	61.75	34.44	58.93	80.60	61.75	34.44	58.93	79.42	59.76	32.95	57.37
	BT-VT	–	–	–	–	80.48	61.76	34.42	58.88	78.61	61.15	30.44	56.73	79.90	61.38	34.27	58.51	68.54	49.27	30.88	49.56
	BV-TT	74.53	47.55	17.72	46.6	80.60	61.76	34.44	58.93	80.60	61.75	34.44	58.93	80.60	61.75	34.44	58.93	79.42	59.76	32.95	57.37
	BV-VT	–	–	–	–	80.48	61.76	34.42	58.88	79.50	60.21	34.39	58.03	79.90	61.38	34.27	58.51	68.54	49.27	30.88	49.56
SAT [8] *(iccv'23)* DeiT-S	BT-TT	79.54	64.59	37.47	60.53	82.92	70.12	52.93	68.65	82.33	67.08	47.01	65.47	82.92	70.13	52.94	68.66	82.92	70.13	52.94	68.66
	BT-VT	–	–	–	–	82.91	69.46	52.82	68.39	79.96	64.78	45.18	63.30	81.19	69.05	52.25	67.49	73.13	57.06	51.02	60.40
	BV-TT	80.09	66.17	43.52	63.26	82.92	70.12	52.93	68.65	82.33	67.08	47.01	65.47	82.92	70.13	52.94	68.66	82.92	70.13	52.94	68.66
	BV-VT	–	–	–	–	82.91	69.46	52.82	68.39	79.96	64.78	45.18	63.30	81.19	69.05	52.25	67.49	73.13	57.06	51.02	60.40

Table S5. Test-set MaxBoxAccV2 that is average of three threshold IoU-30, IoU-50, IoU-70 (here MaxBoxAccV2 is aerage at three threhold) along with the it of WSOL models with different selection criteria on ILSVRC. The *select* column presents (i) BT and BV indicate model selection based on hyperparameter configurations using the test set and validation set, respectively; (ii) TT and VT indicate that the threshold τ is selected using either the test set or validation set. For model selection on the validation set, we consider the GT as a reference, a selection based on the CL performance and the three different pseudo-bboxes generation proposed in this work: RPN, CLIP and SS. Our results for models selected with pseudo-bboxes are comparable to those of GT.

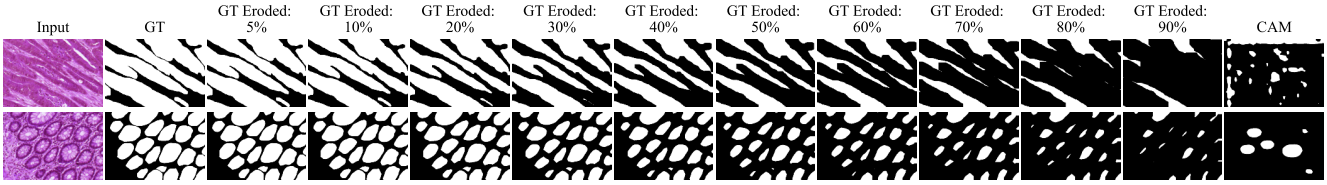

Figure S1. Illustration GT masks, noisy masks at various erosion levels.

we employed two histology image datasets and first simulated noisy masks by applying erosion to the GT masks. This approach allowed us to assess the protocol's robustness against different levels of noise. Additionally for realistic setup, we extracted pseudo-masks from the activation maps of a pre-trained classifier, thereby evaluating the protocol's performance in a realistic setting. Our results indicate that the evaluation protocol consistently maintains performance across varying noise levels, confirming its robustness in both simulated and real-world scenarios.

Datasets. We employed two additional datasets to show the robustness of our proposed protocol; *(i) GLaS dataset* is collected for the diagnosis of colon cancer and comprises 165 images derived from 16 Hematoxylin and Eosin (H&E) stained slides. It includes pixel-level and image-level annotations (benign or malignant). The dataset is divided into 67 images for training, 18 for validation, and 80 for testing. Our evaluation follows the protocol established by [6], where three fully supervised examples per class are used for best model over LOC. *(ii) CAMELYON dataset* is a patch-based benchmark extracted using Camelyon16 dataset, which consists of 399 whole slide images with two classes (normal and metastatic) used for detecting metastases in H&E-stained tissue sections of sentinel lymph nodes from breast cancer patients. Following the extraction protocol outlined in [6], patches of size 512×512 are annotated at both image and pixel levels. This dataset comprises a total of 48,870 images, with 24,348 for training, 8,850 for validation, and 15,664 for testing. From the validation set, six fully supervised examples per class are randomly selected to determine best model, as suggested by [6].

Evaluation measures. The GLaS and CAMELYON datasets provide pixel-wise annotations (masks) rather than bboxes, necessitating the use of pixel average precision (PxAP) [3, 6] to evaluate localization accuracy. Following the standard WSOL pipeline, we first employ min-max normalization on these activation maps and apply various thresholds to the activation maps for producing localization maps.

Generation of noisy and pseudo-GT masks. The objective of this study is to evaluate the robustness of the proposed model selection techniques under conditions of noisy masks. To this end, we perturb the GT masks through erosion, simulating scenarios with noisy or imprecise GT annotations commonly encountered in practical applications. The noisy GT masks are generated by applying erosion with varying filter sizes and iterations until the masks are degraded to a specified extent, corresponding to eleven predefined noise levels, as detailed in Tab.S6 and examples of noisy masks are illustrated in Fig.S1.

Results. Tab.S6 compares the performance of models (ResNet50) trained with pseudo-masks and generated pseudo-masks against those trained with GT masks, demonstrating the robustness of our evaluation protocol across varying noise levels. The results include localization performance for models using pseudo-masks generated by a pre-trained InceptionV3 classifier, an architecture distinct from the one used for training our models. Despite significant increases in mask error across different noise levels on both the GLaS and CAMELYON datasets, localization performance remains relatively stable, even at the highest noise level. Additionally, pseudo-masks generated from a different architecture show reasonable localization accuracy, further validating the generalizability of our evaluation protocol to diverse types of annotation noise, and underscoring its robustness in practical settings.

Pseudo Mask	GLaS		CAMELYON	
	Mask Error 1-AUC	LOC PxAP	Mask Error 1-AUC	LOC PxAP
GT-Mask	0.0	69.78	0.0	30.55
Perturbed GT-Mask with erosion level:				
5%	1.86	70.09	1.64	29.74
10%	4.30	70.3	4.23	38.87
20%	9.22	70.40	9.29	29.74
30%	14.61	70.02	14.26	30.55
40%	19.15	69.83	19.17	29.74
50%	24.34	70.58	24.22	29.74
60%	29.18	67.68	29.14	29.74
70%	34.16	68.38	34.11	29.74
80%	39.21	67.98	39.17	29.74
90%	44.12	68.65	44.14	29.74
CAM [12] Pseudo-Mask	46.54	66.41	28.10	29.74

Table S6. Impact of varying levels of erosion in GT masks on LOC performance (PxAP) compared to GT-mask and pseudo-masks. It shows that despite substantial increases in mask error (1-AUC) across different noise levels, the LOC performance remains stable, particularly when using pseudo-masks, highlighting the robustness and generalizability of our evaluation protocol across different domains.

References

[1] H. Bai, R. Zhang, J. Wang, and X. Wan. Weakly supervised object localization via transformer with implicit spatial cali-

bration. *ECCV*, 2022.

[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] J. Choe, S. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, 2020.

[4] J. Choe and H. Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, 2019.

[5] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *ICCV*, 2021.

[6] J. Rony, S. Belharbi, J. Dolz, I. Ben Ayed, L. McCaffrey, and E. Granger. Deep weakly-supervised learning methods for classification and localization in histology images: A survey. *MLBI*, 2:96–150, 2023.

[7] K. Kumar Singh and Y. Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017.

[8] P. Wu, W. Zhai, Y. Cao, J. Luo, and ZJ. Zha. Spatial-aware token for weakly supervised object localization. In *ICCV*, 2023.

[9] S. Yang, Y. Kim, Y. Kim, and C. Kim. Combinational class activation maps for weakly supervised object localization. In *WACV*, 2020.

[10] J. Zhang, S. Adel Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 126(10):1084–1102, 2018.

[11] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018.

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