Evaluating Weakly Supervised Object Localization Methods Right

Junsuk Choe[1,3*] Seong Joon Oh[2*] Seungho Lee[1]
Sanghyuk Chun[3] Zeynep Akata[4] Hyunjung Shim[1†]

[1]School of Integrated Technology, Yonsei University
[2]Clova AI Research, LINE Plus Corp.
[3]Clova AI Research, NAVER Corp.
[4]University of Tuebingen

Abstract

Weakly-supervised object localization (WSOL) has gained popularity over the last years for its promise to train localization models with only image-level labels. Since the seminal WSOL work of class activation mapping (CAM), the field has focused on how to expand the attention regions to cover objects more broadly and localize them better. However, these strategies rely on full localization supervision to validate hyperparameters and for model selection, which is in principle prohibited under the WSOL setup. In this paper, we argue that WSOL task is ill-posed with only image-level labels, and propose a new evaluation protocol where full supervision is limited to only a small held-out set not overlapping with the test set. We observe that, under our protocol, the five most recent WSOL methods have not made a major improvement over the CAM baseline. Moreover, we report that existing WSOL methods have not reached the few-shot learning baseline, where the full-supervision at validation time is used for model training instead. Based on our findings, we discuss some future directions for WSOL. Source code and dataset are available at https://github.com/clovaai/wsolevaluation.

1. Introduction

As human labeling for every object is too costly and weakly-supervised object localization (WSOL) requires *only* image-level labels, the WSOL research has gained significant momentum [58, 56, 57, 6, 25, 55] recently.

Among these, class activation mapping (CAM) [58] uses the intermediate classifier activations focusing on the most discriminative parts of the objects to localize the objects of the target class. As the aim in object localization is to cover the full extent of the object, focusing only on the most discriminative parts of the objects is a limitation. WSOL techniques since CAM have focused on this limitation and have proposed different architectural [56, 57, 6] and data-augmentation [25, 55] solutions. The reported state-of-the-art WSOL performances have made a significant improvement over the CAM baseline, from 49.4% to 62.3% [6] and 43.6% to 48.7% [6] top-1 localization performances on Caltech-UCSD Birds-200-2011 [52] and ImageNet [41], respectively. However, these techniques have introduced a set of hyperparameters for suppressing the discriminative cues of CAM and different ways for selecting these hyperparameters. One of such hyperparameters is the operating threshold τ for generating object bounding boxes from the score maps. Among others, the mixed policies for selecting τ has contributed to the illusory improvement of WSOL performances over the years; see Figure 1.

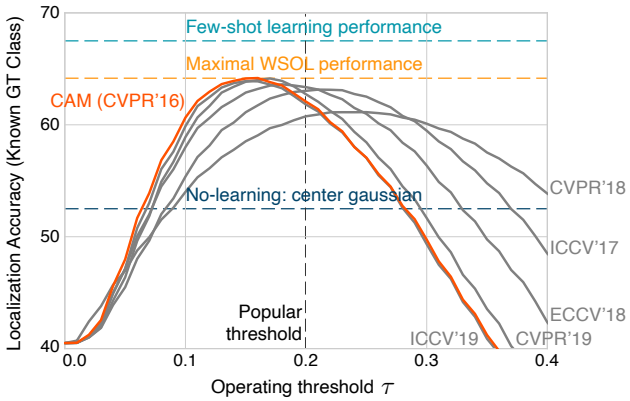


Figure 1. **WSOL 2016-2019.** Recent improvements in WSOL are illusory due to (1) different amount of implicit full supervision through validation and (2) a fixed score-map threshold (usually $\tau = 0.2$) to generate object boxes. Under our evaluation protocol with the same validation set sizes and oracle τ for each method, CAM is still the best. In fact, our few-shot learning baseline, *i.e.* using the validation supervision (10 samples/class) at training time, outperforms existing WSOL methods. Results on ImageNet.

Due to the lack of a unified definition of the WSOL task, we revisit the problem formulation of WSOL and show that WSOL problem is ill-posed in general without any localiza-

*Equal contribution. Work done at Clova AI Research.
†Hyunjung Shim is a corresponding author.

tion supervision. Towards a well-posed setup, we propose a new WSOL setting where a small held-out set with full supervision is available to the learners.

Our contributions are as follows. (1) Propose new experimental protocol that uses a fixed amount of full supervision for hyperparameter search and carefully analyze six WSOL methods on three architectures and three datasets. (2) Propose new evaluation metrics as well as data, annotations, and benchmarks for the WSOL task at https://github.com/clovaai/wsolevaluation. (3) Show that WSOL has not progressed significantly since CAM, when the calibration dependency and the different amounts of full supervision are factored out. Moreover, searching hyperparameters on a held-out set consisting of 5 to 10 full localization supervision per class often leads to significantly lower performance compared to the few-shot learning (FSL) baselines that use the full supervision directly for model training. Finally, we suggest a shift of focus in future WSOL research: consideration of learning paradigms utilizing both weak and full supervisions, and other options for resolving the ill-posedness of WSOL (*e.g.* background-class images).

2. Related Work

By model output. Given an input image, *semantic segmentation* models generate pixel-wise class predictions [11, 31], *object detection* models [11, 13] output a set of bounding boxes with class predictions, and *instance segmentation* models [27, 7, 18] predict a set of disjoint masks with class *and* instance labels. *Object localization* [41], on the other hand, assumes that the image contains an object of single class and produces a binary mask or a bounding box around that object coming from the class of interest.

By type of supervision. Since bounding box and mask labels cost significantly more than image-level labels, *e.g.* categories [2], researchers have considered different types of localization supervision: image-level labels [35], gaze [34], points [2], scribbles [26], boxes [8], or a mixture of multiple types [21]. Our work is concerned with the object localization task with only image-level category labels [33, 58].

By amount of supervision. Learning from a small amount of labeled samples per class is referred to as few-shot learning (FSL) [53]. We recognize the relationship between our new WSOL setup and the FSL paradigm; we consider FSL methods as baselines for future WSOL methods.

WSOL works. Class activation mapping (CAM) [58] turns a fully-convolutional classifier into a score map predictor by considering the activations before the global average pooling layer. Vanilla CAM has been criticized for its focus on the small discriminative part of the object. Researchers have considered dropping regions in inputs at random [25, 55] to diversify the cues used for recognition. Adversarial erasing techniques [56, 6] drop the most discriminative part at the current iteration. Self-produced guidance (SPG) [57] is trained with auxiliary foreground-background masks generated by its own activations. Other than object classification in static images, there exists work on localizing informative video frames for action recognition [37, 28, 54], but they are beyond the scope of our analysis.

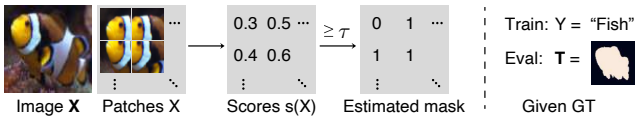


Figure 2. **WSOL as MIL.** WSOL is interpreted as a patch classification task trained with multiple-instance learning (MIL). The score map $s(\mathbf{X})$ is thresholded at τ to estimate the mask $\mathbf{T}$.

Relation to explainability. WSOL methods share similarities with the model explainability [17], specifically the *input attribution* task: analyzing which pixels have led to the image classification results [16]. There are largely two streams of work on visual input attribution: variants of input gradients [50, 45, 42, 48, 43, 49, 24, 36] and counterfactual reasoning [39, 12, 59, 40, 15, 20]. While they can be viewed as WSOL methods, we have not included them in our studies because they are seldom evaluated in WSOL benchmarks. Analyzing their possibility as WSOL methods is an interesting future study.

Our scope. We study the WSOL task, rather than weakly-supervised detection, segmentation, or instance segmentation. The terminologies tend to be mixed in the earlier works of weakly-supervised learning [44, 14, 9, 47]. Extending our analysis to other weakly-supervised learning tasks is valid and will be a good contribution to the respective communities.

3. Problem Formulation of WSOL

We define and formulate the weakly-supervised object localization (WSOL) task as an image patch classification and show the ill-posedness of the problem. We will discuss possible modifications to resolve the ill-posedness in theory.

3.1. WSOL Task as Multiple Instance Learning

Given an image $\mathbf{X} \in \mathbb{R}^{H \times W}$, **object localization** is the task to identify whether or not the pixel belongs to the object of interest, represented via dense binary mask $\mathbf{T} = (T_{11}, \cdots, T_{HW})$ where $T_{ij} \in \{0, 1\}$ and (i, j) indicate the pixel indices. When the training set consists of precise image-mask pairs $(\mathbf{X}, \mathbf{T})$, we refer to the task as **fully-supervised object localization (FSOL)**. In this paper, we consider the case when only an image-level label $Y \in \{0, 1\}$ for global presence of the object of interest is provided per training image $\mathbf{X}$. This task is referred to as the **weakly-supervised object localization (WSOL)**.

One can treat an input image $\mathbf{X}$ as a bag of stride-1 sliding window patches of suitable side lengths, h and w:

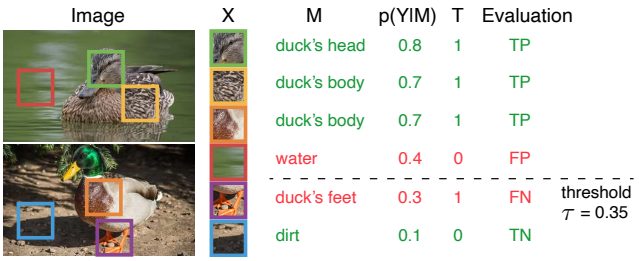


Figure 3. **Ill-posed WSOL: An example.** Even the true posterior $s(M) = p(Y|M)$ may not lead to the correct prediction of T if background cues are more associated with the class than the foreground cues (*e.g.* $p(\text{duck}|\text{water}) > p(\text{duck}|\text{feet})$).

$(X_{11}, \cdots, X_{HW})$ with $X_{ij} \in \mathbb{R}^{h \times w}$. The object localization task is then the problem of predicting the object presence T_{ij} at the image patch X_{ij}. The weak supervision imposes the requirement that each training image $\mathbf{X}$, represented as $(X_{11}, \cdots, X_{HW})$, is only collectively labeled with a single label $Y \in \{0, 1\}$ indicating whether at least one of the patches represents the object. This formulation is an example of the multiple-instance learning (MIL) [22], as observed by many traditional WSOL works [35, 44, 14, 47].

Following the patch classification point of view, we formulate WSOL task as a mapping from patches X to the binary labels T (indices dropped). We assume that the patches X, image-level labels Y, and the pixel-wise labeling T in our data arise in an i.i.d. fashion from the joint distribution $p(X, Y, T)$. See Figure 2 for an overview. The aim of WSOL is to produce a scoring function $s(X)$ such that thresholding it at τ closely approximates binary label T. Many existing approaches for WSOL, including CAM [58], use the scoring rules based on the posterior $s(X) = p(Y|X)$. See Appendix §A.1 for the interpretation of CAM as pixel-wise posterior approximation.

3.2. When is WSOL ill-posed?

We show that if background cues are more strongly associated with the target labels T than some foreground cues, the localization task cannot be solved, even when we know the exact posterior $p(Y|X)$ for the image-level label Y. We will make some strong assumptions in favor of the learner, and then show that WSOL still cannot be perfectly solved.

We assume that there exists a finite set of **cue** labels $\mathcal{M}$ containing all patch-level concepts in natural images. For example, patches from a duck image are one of {duck's head, duck's feet, sky, water, $\cdots$} (see Figure 3). We further assume that every patch X is equivalently represented by its cue label $M(X) \in \mathcal{M}$. Therefore, from now on, we write M instead of X in equations and examine the association arising in the joint distribution $p(M, Y, T)$. We write $M^{\text{fg}}, M^{\text{bg}} \in \mathcal{M}$ for foreground and background cues.

We argue that, even with access to the joint distribution $p(Y, M)$, it may not be possible to make perfect predictions for the patch-wise labels $T(M)$ (proof in Appendix §A.2).

Lemma 3.1. *Assume that the true posterior $p(Y|M)$ with a continuous pdf is used as the scoring rule $s(M) = p(Y|M)$. Then, there exists a scalar $\tau \in \mathbb{R}$ such that $s(M) \geq \tau$ is identical to T if and only if the foreground-background posterior ratio $\frac{p(Y=1|M^{fg})}{p(Y=1|M^{bf})} \geq 1$ almost surely, conditionally on the event $\{T(M^{fg}) = 1$ and $T(M^{bf}) = 0\}$.*

In other words, if the posterior likelihood for the image-level label Y given a foreground cue M^{fg} is less than the posterior likelihood given background M^{bg} for some foreground and background cues, no WSOL method can make a correct prediction. This pathological scenario is described in Figure 3: Duck's feet are less seen in duck images than the water background. Such cases are abundant in user-collected data (Appendix Figure 1).

This observation implies a data-centric solution towards well-posed WSOL: we can augment (1) positive samples ($Y = 1$) with more less-represented foreground cues (*e.g.* duck images with feet) and (2) negative samples ($Y = 0$) with more target-correlated background cues (*e.g.* non-duck images with water background). Such data-centric approaches are promising future directions for WSOL.

How have WSOL methods addressed the ill-posedness? Previous solutions to the WSOL problem have sought architectural modifications [56, 57, 6] and data augmentation [25, 55] schemes that typically require heavy hyperparameter search and model selection, which are a form of implicit localization supervision. For example, [25] has found the operating threshold τ via "observing a few qualitative results", while others have evaluated their models over the test set to select reasonable hyperparameter values (Table 1 of [25], Table 6 of [56], and Table 1 of [6]). [57] has performed a "grid search" over possible values. We argue that certain level of localization labels are inevitable for WSOL. In the next section, we propose to allow a fixed number of fully labeled samples for hyperparameter search and model selection for a more realistic evaluation.

4. Evaluation Protocol for WSOL

We reformulate the WSOL evaluation based on our observation of the ill-posedness. We define performance metrics, benchmarks, and the hyperparameter search procedure.

4.1. Evaluation metrics

The aim of WSOL is to produce score maps, where their pixel value s_{ij} is higher on foreground $T_{ij} = 1$ and lower on background $T_{ij} = 0$ (§3.1). We discuss how to quantify the above conditions and how prior evaluation metrics have failed to clearly measure the relevant performance. We then propose the `MaxBoxAcc` and `PxAP` metrics for bounding box and mask ground truths, respectively.

The *localization accuracy* [41] metric entangles classification and localization performances by counting the num-

ber of images where both tasks are performed correctly. We advocate the measurement of localization performance alone, as the goal of WSOL is to localize objects (§3.1) and not to classify images correctly. To this end, we only consider the score maps s_{ij} corresponding to the ground-truth classes in our analysis. Metrics based on such are commonly referred to as the *GT-known* metrics [25, 56, 57, 6].

A common practice in WSOL is to normalize the score maps per image because the score statistics differ vastly across images. Either max normalization (divide through by $\max_{ij} s_{ij}$) or min-max normalization (additionally map $\min_{ij} s_{ij}$ to zero) has been used; see Appendix §B.1 for the full summary. We always use the min-max normalization.

After normalization, WSOL methods threshold the score map at τ to generate a tight box around the binary mask $\{(i,j) \mid s_{ij} \geq \tau\}$. WSOL metrics then measure the quality of the boxes. τ is typically treated as a fixed value [58, 56, 55] or a hyperparameter to be tuned [25, 57, 6]. We argue that the former is misleading because the ideal threshold τ depends heavily on the data and model architecture and fixing its value may be disadvantageous for certain methods. To fix the issue, we propose new evaluation metrics that are independent of the threshold τ.

Masks: `PxAP`. When masks are available for evaluation, we measure the pixel-wise precision and recall [1]. Unlike single-number measures like mask-wise IoU, those metrics allow users to choose the preferred operating threshold τ that provides the best precision-recall trade-off for their downstream applications. We define the **pixel precision and recall at threshold** τ as:

$$\texttt{PxPrec}(\tau) = \frac{|\{s_{ij}^{(n)} \geq \tau\} \cap \{T_{ij}^{(n)} = 1\}|}{|\{s_{ij}^{(n)} \geq \tau\}|} \quad (1)$$

$$\texttt{PxRec}(\tau) = \frac{|\{s_{ij}^{(n)} \geq \tau\} \cap \{T_{ij}^{(n)} = 1\}|}{|\{T_{ij}^{(n)} = 1\}|} \quad (2)$$

For threshold independence, we define and use the **pixel average precision**, $\texttt{PxAP} := \sum_l \texttt{PxPrec}(\tau_l)(\texttt{PxRec}(\tau_l) - \texttt{PxRec}(\tau_{l-1}))$, the area under curve of the pixel precision-recall curve. We use `PxAP` as the final metric in this paper.

Bounding boxes: `MaxBoxAcc`. Pixel-wise masks are expensive to collect; many datasets only provide box annotations. Since it is not possible to measure exact pixel-wise precision and recall with bounding boxes, we suggest a surrogate in this case. Given the ground truth box B, we define the **box accuracy at score map threshold** τ **and IoU threshold** δ, $\texttt{BoxAcc}\ (\tau, \delta)$ [58, 41], as:

$$\texttt{BoxAcc}(\tau, \delta) = \frac{1}{N} \sum_n 1_{\text{IoU}\left(\text{box}(s(\mathbf{X}^{(n)}), \tau), B^{(n)}\right) \geq \delta} \quad (3)$$

where $\text{box}(s(\mathbf{X}^{(n)}), \tau)$ is the tightest box around the largest-area connected component of the mask $\{(i,j) \mid s(X_{ij}^{(n)}) \geq \tau\}$. In datasets where more than one bounding box are provided (*e.g.* ImageNet), we count the number of images where the box prediction overlaps with *at least one* of the ground truth boxes with IoU $\geq \delta$. When δ is 0.5, the metric is identical to the commonly-called *GT-known localization accuracy* [25] or *CorLoc* [10], but we suggest a new naming to more precisely represent what is being measured. For score map threshold independence, we report the box accuracy at the optimal threshold τ, the **maximal box accuracy** $\texttt{MaxBoxAcc}(\delta) := \max_\tau \texttt{BoxAcc}(\tau, \delta)$, as the final performance metric. In this paper, we set δ to 0.5, following the prior works [58, 25, 56, 57, 6, 55].

Statistics	ImageNet	CUB	OpenImages
#Classes	1000	200	100
#images/class			
`train-weaksup`	∼1.2K	∼30	∼300
`train-fullsup`	10	∼5	25
`test`	10	∼29	50

Table 1. **Dataset statistics.** "∼" indicates that the number of images per class varies across classes and the average value is shown.

Better box evaluation: `MaxBoxAccV2`. After the acceptance at CVPR 2020, we have developed an improved version of `MaxBoxAcc`. It is better in two aspects. (1) `MaxBoxAcc` measures the performance at a fixed IoU threshold ($\delta = 0.5$), only considering a specific level of fineness of localization outputs. We suggest averaging the performance across $\delta \in \{0.3, 0.5, 0.7\}$ to address diverse demands for localization fineness. (2) `MaxBoxAcc` takes the *largest* connected component for estimating the box, assuming that the object of interest is usually large. We remove this assumption by considering the best match between the set of all estimated boxes and the set of all ground truth boxes. We call this new metric as `MaxBoxAccV2`. For future WSOL researches, we encourage using the `MaxBoxAccV2` metric. The code is already available in our repository. We show the evaluation results under the new metric in the in Appendix §C.7.

4.2. Data splits and hyperparameter search

For a fair comparison of the WSOL methods, we fix the amount of full supervision for hyperparameter search. As shown in Table 1 we propose three disjoint splits for every dataset: `train-weaksup`, `train-fullsup`, and `test`. The `train-weaksup` contains images with weak supervision (the image-level labels). The `train-fullsup` contains images with full supervision (either bounding box or binary mask). It is left as freedom for the user to utilize it for hyperparameter search, model selection, ablative studies, or even model fitting. The `test` split contains images with full supervision; it must be used only for the final performance report. For example, checking the `test` results multiple times with different model configurations violates the protocol as the learner implicitly uses more full supervision than allowed.

As WSOL benchmark datasets, ImageNet [41] and

Caltech-UCSD Birds-200-2011 (CUB) [52] have been extensively used. For ImageNet, the 1.2M "train" and 10K "validation" images for 1 000 classes are treated as our `train-weaksup` and `test`, respectively. For `train-fullsup`, we use the ImageNetV2 [38]. We have annotated bounding boxes on those images. CUB has 5 994 "train" and 5 794 "test" images for 200 classes. We treat them as our `train-weaksup` and `test`, respectively. For `train-fullsup`, we have collected 1 000 extra images (~ 5 images per class) from Flickr, on which we have annotated bounding boxes. For ImageNet and CUB we use the oracle box accuracy `BoxAcc`.

We contribute a new WSOL benchmark based on the OpenImages instance segmentation subset [3]. It provides a fresh WSOL benchmark to which the models have not yet overfitted. To balance the original OpenImages dataset, we have sub-sampled 100 classes and have randomly selected 29 819, 2 500, and 5 000 images from the original "train", "validation", and "test" splits as our `train-weaksup`, `train-fullsup`, and `test` splits, respectively. We use the pixel average precision `PxAP`. A summary of dataset statistics is in Table 1. Details on data collection and preparation are in Appendix §B.2.

Hyperparameter search. To make sure that the same amount of localization supervision is provided for each WSOL method, we refrain from employing any source of human prior outside the `train-fullsup` split. If the optimal hyperparameter for an arbitrary dataset and architecture is not available by default, we subject it to the hyperparameter search algorithm. For each hyperparameter, its *feasible range*, as opposed to *sensible range*, is used as the search space, to minimize the impact of human bias.

We employ the random search hyperparameter optimization [4]; it is simple, effective, and parallelizable. For each WSOL method, we sample 30 hyperparameters to train models on `train-weaksup` and validate on `train-fullsup`. The best hyperparameter combination is then selected. Since running 30 training sessions is costly for ImageNet (1.2M training images), we use 10% of images in each class for fitting models during the search. We verify in Appendix §B.4 that the ranking of hyperparameters is preserved even if the training set is sub-sampled.

5. Experiments

5.1. Evaluated Methods

We evaluate six widely used WSOL methods published in peer-reviewed venues. We describe each method in chronological order and discuss the set of hyperparameters. The full list of hyperparameters is in Appendix §C.1.

Class activation mapping (CAM) [58] trains a classifier of fully-convolutional backbone with the global average pooling (GAP) structure. At test time, CAM uses the logit outputs before GAP as the score map s_{ij}. CAM has the learning rate and the score-map resolution as hyperparameters and all five methods below use CAM in the background.

Hide-and-seek (HaS) [25] is a data augmentation technique that randomly selects grid patches to be dropped. The hyperparameters are the drop rate and grid size.

Adversarial complementary learning (ACoL) [56] proposes an architectural solution: a two-head architecture where one adversarially erases the high-scoring activations in the other. The erasing threshold is a hyperparameter.

Self-produced guidance (SPG) [57] is another architectural solution where internal pseudo-pixel-wise supervision is synthesized on the fly. Three tertiary pixel-wise masks (foreground, unsure, background) are generated from three different layers using two thresholding hyperparameters for each mask and are used as auxiliary supervisions.

Attention-based dropout layer (ADL) [6] has proposed a module that, like ACoL, adversarially produces drop masks at high-scoring regions, while not requiring an additional head. Drop rate and threshold are the hyperparameters.

CutMix [55] is a data augmentation technique, where patches in training images are cut and pasted to other images during training. The target labels are also mixed. The hyperparameters are the size prior α and the mix rate r.

Few-shot learning (FSL) baseline. The full supervision in `train-fullsup` used for validating WSOL hyperparameters can be used for training a model itself. Since only a few fully labeled samples per class are available, we refer to this setting as the few-shot learning (FSL) baseline.

As a simple baseline, we consider a foreground saliency mask predictor [29]. We alter the last layer of a fully convolutional network (FCN) into a 1×1 convolutional layer with $H \times W$ score map output. Each pixel is trained with the binary cross-entropy loss against the target mask, as done in [5, 31, 32]. For OpenImages, the pixel-wise masks are used as targets; for ImageNet and CUB, we build the mask targets by labeling pixels inside the ground truth boxes as foreground [23]. At inference phase, the $H \times W$ score maps are evaluated with the box or mask metrics.

Center-gaussian baseline. The Center-gaussian baseline generates isotropic Gaussian score maps centered at the images. We set the standard deviation to 1, but note that it does not affect the `MaxBoxAcc` and `PxAP` measures. This provides a no-learning baseline for every localization method.

5.2. Comparison of WSOL methods

We evaluate the six WSOL methods over three backbone architectures, *i.e.* VGG-GAP [46, 58], InceptionV3 [51], and ResNet50 [19], and three datasets, *i.e.* CUB, ImageNet and OpenImages. For each (method, backbone, dataset) tuple, we have randomly searched the optimal hyperparameters over the `train-fullsup` with 30 trials, totalling about 9 000 GPU hours. Since the sessions are paralleliz-

Methods	ImageNet (MaxBoxAcc) VGG	Inception	ResNet	Mean	CUB (MaxBoxAcc) VGG	Inception	ResNet	Mean	OpenImages (PxAP) VGG	Inception	ResNet	Mean	Total Mean
CAM [58]	61.1	65.3	64.2	63.5	71.1	62.1	73.2	68.8	58.1	61.4	58.0	59.1	63.8
HaS [25]	+0.7	+0.1	-1.0	-0.1	+5.2	-4.4	+4.9	+1.9	-1.2	-2.9	+0.2	-1.3	+0.2
ACoL [56]	-0.8	-0.7	-2.5	-1.4	+1.2	-2.5	-0.5	-0.6	-3.4	+1.6	-0.2	-0.7	-0.9
SPG [57]	+0.5	+0.1	-0.7	+0.0	-7.4	+0.7	-1.8	-2.8	-2.2	+1.0	-0.3	-0.5	-1.1
ADL [6]	-0.3	-3.8	+0.0	-1.4	+4.6	+1.3	+0.3	+2.0	+0.2	+0.7	-3.7	-0.9	-0.1
CutMix [55]	+1.0	+0.1	-0.3	+0.3	+0.8	+3.4	-5.4	-0.4	+0.1	+0.3	+0.7	+0.4	+0.1
Best WSOL	62.2	65.5	64.2	63.8	76.2	65.5	78.1	70.8	58.3	63.0	58.6	59.5	64.0
FSL baseline	62.8	68.7	67.5	66.3	86.3	94.0	95.8	92.0	61.5	70.3	74.4	68.7	75.7
Center baseline	52.5	52.5	52.5	52.5	59.7	59.7	59.7	59.7	45.8	45.8	45.8	45.8	52.3

Table 2. **Re-evaluating WSOL.** How much have WSOL methods improved upon the vanilla CAM model? `test` split results are shown, relative to the vanilla CAM performance (increase or decrease). Hyperparameters have been optimized over the identical `train-fullsup` split for all WSOL methods and the FSL baseline: (10,5,5) full supervision/class for (ImageNet,CUB,OpenImages). Reported results are in the Appendix Table 5; classification accuracies are in Appendix Table 4.

able, it has taken only about 200 hours over 50 P40 GPUs to obtain the results. The results are shown in Table 2. We use the same batch sizes and training epochs to enforce the same computational budget. The checkpoints that achieves the best localization performance on `train-fullsup` are used for evaluation.

Contrary to the improvements reported in prior work (Appendix Table 5), recent WSOL methods have not led to major improvements compared to CAM, when validated in the same data splits and same evaluation metrics. On ImageNet, methods after CAM are generally struggling: only CutMix has seen a boost of +0.3pp on average. On CUB, ADL has attained a +2.0pp gain on average, but ADL fails to work well on other benchmarks. On the new WSOL benchmark, OpenImages, no method has improved over CAM, except for CutMix (+0.4pp on average). The best overall improvements over CAM (63.8% total mean) is a mere +0.2pp boost by HaS. In general, we observe a random mixture of increases and decreases in performance over the baseline CAM, depending on the architecture and dataset. An important result in the table to be discussed later is the comparison against the few-shot learning baseline (§5.5).

Some reasons for the discrepancy between our results and the reported results include (1) the confounding of the actual score map improvement and the calibration scheme, (2) different types and amounts of full supervision employed under the hood, and (3) the use of different training settings (*e.g.* batch size, learning rates, epochs). More details about the training settings are in Appendix §C.3.

Which checkpoint is suitable for evaluation? After the acceptance by CVPR 2020, we believe that it is inappropriate to use the best checkpoint for WSOL evaluation. This is because the best localization performances are achieved before convergence in many cases (Appendix §C.2. At early epochs, the localization performance fluctuates a lot, so the peak performance is noise rather than the real performance.

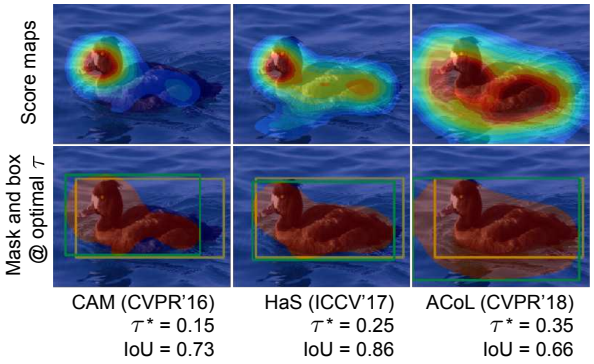


Figure 4. **Selecting τ.** Measuring performance at a fixed threshold τ can lead to a false sense of improvement. Compared to CAM, HaS and ACoL expand the score maps, but they do not necessarily improve the box qualities (IoU) at the optimal $\tau^\star$. Predicted and ground-truth boxes are shown as green and yellow boxes.

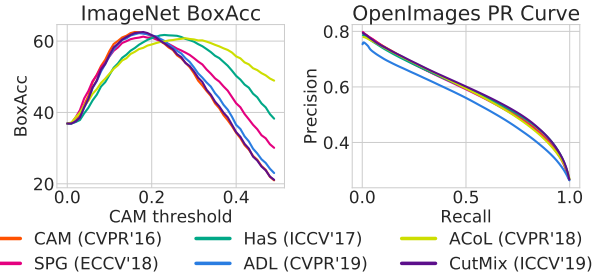
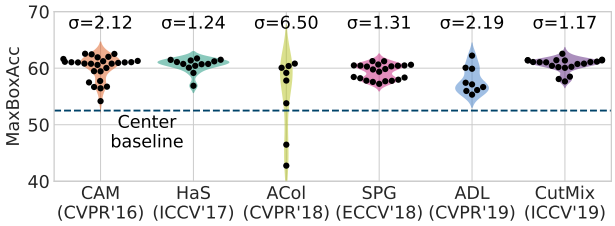


Figure 5. **Performance at varying operating thresholds.** ImageNet: `BoxAcc`(τ) versus τ. OpenImages: `PxPrec`(τ) versus `PxRec`(τ). Both use ResNet.

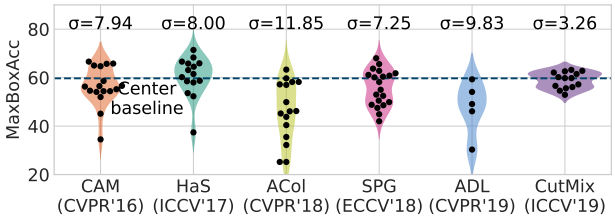
Hence, we recommend using the final checkpoint for future WSOL researchers. The evaluation results are shown in Appendix Table 6.

5.3. Score calibration and thresholding

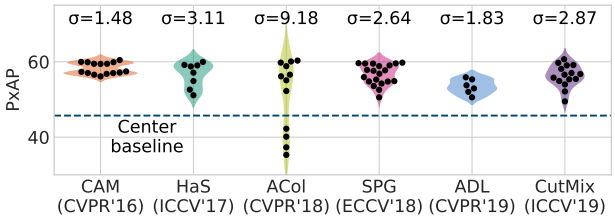
WSOL evaluation must focus more on score map evaluation, independent of the calibration. As shown in Figure 4 the min-max normalized score map for CAM predicts



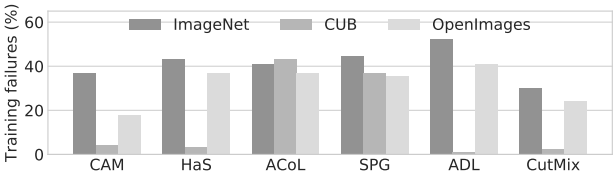
(a) ResNet50 architecture, ImageNet dataset.



(b) ResNet50 architecture, CUB dataset.



(c) ResNet50 architecture, OpenImages dataset.



(d) Ratio of training failures in different WSOL methods.

Figure 6. **Results of the 30 hyperparameter trials.** ImageNet performances of all 30 randomly chosen hyperparameter combinations for each method, with ResNet50 backbone. The violin plots show the estimated distributions (kernel density estimation) of performances. σ are the sample standard deviations.

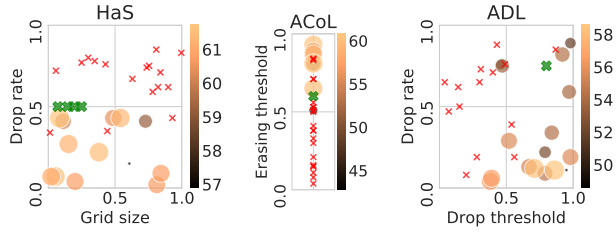


Figure 7. **Impact of hyperparameters for feature erasing.** Color and size of the circles indicate the performance at the corresponding hyperparameters. ×: non-convergent training sessions. ✖: hyperparameters suggested by the original papers.

a peaky foreground score on the duck face, While HaS and ACoL score maps show more distributed scores in body areas, demonstrating the effects of adversarial erasing during training. However, the maximal IoU performances do not differ as much. This is because WSOL methods exhibit different score distributions (Figure 5 and Appendix §C.4). Fixing the operating threshold τ at a pre-defined value, therefore, can lead to an apparent increase in performance without improving the score maps.

Under our threshold-independent performance measures (`MaxBoxAcc` and `PxAP`) shown in Figure 5, we observe that (1) the methods have different optimal $\tau^\star$ on ImageNet and (2) the methods do not exhibit significantly different `MaxBoxAcc` or `PxAP` performances. This provides an explanation of the lack of improvement observed in Table 2. We advise future WSOL researchers to report the threshold-independent metrics.

5.4. Hyperparameter analysis

Different types and amounts of full supervision used in WSOL methods manifest in the form of model hyperparameter selection (§3). Here, we measure the impact of the validation on `train-fullsup` by observing the performance distribution among 30 trials of random hyperparameters. We then study the effects of feature-erasing hyperparameters, a common hyperparameter type in WSOL methods.

Performance with 30 hyperparameter trials. To measure the sensitivity of each method to hyperparameter choices, we plot the performance distribution of the intermediate models in the 30 random search trials. We say that a training session is *non-convergent* if the training loss is larger than 2.0 at the last epoch. We show the performance distributions of the converged sessions, and report the ratio of non-convergent sessions separately.

Our results in Figure 6 indicate the diverse range of performances depending on the hyperparameter choice. Vanilla CAM is among the less sensitive, with the smallest standard deviation $\sigma = 1.5$ on OpenImages. This is the natural consequence of its minimal use of hyperparameters. We thus suggest to use the vanilla CAM when absolutely no full supervision is available. ACoL and ADL tend to have greater variances across benchmarks ($\sigma = 11.9$ and 9.8 on CUB). We conjecture that the drop threshold for adversarial erasing is a sensitive hyperparameter.

WSOL on CUB are generally struggling: random hyperparameters often show worse performance than the center baseline (66% cases). We conjecture that CUB is a disadvantageous setup for WSOL: as all images contain birds, the models only attend on bird parts for making predictions. We believe adding more non-bird images can improve the overall performances (§3.2).

We show the non-convergence statistics in Figure 6d. Vanilla CAM exhibit a stable training: non-convergence rates are lowest on OpenImages and second lowest on ImageNet. ACoL and SPG suffer from many training failures, especially on CUB (43% and 37%, respectively).

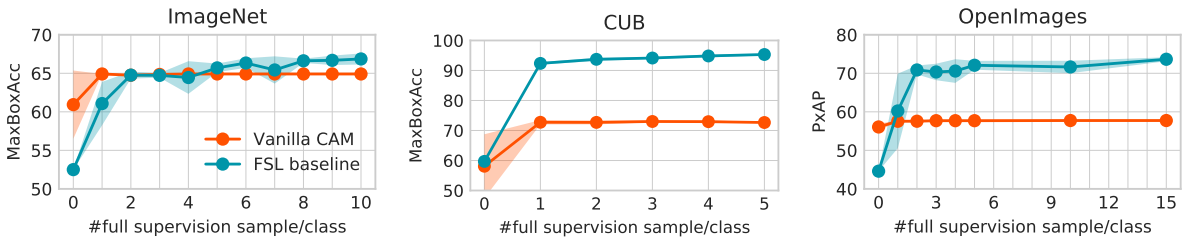


Figure 8. **WSOL versus few-shot learning.** The mean and standard error of models trained on three samples of full-supervision subsets are reported. ResNet50 is used throughout. At 0 full supervision, Vanilla CAM=random-hyperparameter and FSL=center-gaussian baseline.

In conclusion, vanilla CAM is stable and robust to hyperparameters. Complicated design choices introduced by later methods only seem to lower the overall performances rather than providing new avenues for performance boost.

Effects of erasing hyperparameters. Many WSOL methods since CAM have introduced different forms of erasing to encourage models to extract cues from broader regions (§5.1). We study the contribution of such hyperparameters in ADL, HaS, and ACoL in Figure 7. We observe that the performance improves with higher erasing thresholds (ADL drop threshold and ACoL erasing threshold). We also observe that lower drop rates leads to better performances (ADL and HaS). The erasing hyperparameters introduced since CAM only negatively impact the performance.

5.5. Few-shot learning baselines

Given that WSOL methods inevitably utilize some form of full localization supervision (§3), it is important to compare them against the few-shot learning (FSL) baselines that use it for model tuning itself.

Performances of the FSL baselines (§4.2) are presented in Table 2. Our simple FSL method performs better than the vanilla CAM at 10, 5, and 5 fully labeled samples per class for ImageNet, CUB, and OpenImages, respectively. The mean FSL accuracy on CUB is 92.0%, which is far better than that of the maximal WSOL performance of 70.8%.

We compare FSL against CAM at different sizes of `train-fullsup` in Figure 8. We simulate the zero-fully-labeled WSOL performance with a set of randomly chosen hyperparameters (§5.4); for FSL, we simulate the no-learning performance through the center-gaussian baseline.

FSL baselines surpass the CAM results already at 1-2 full supervision per class for CUB and OpenImages (92.4 and 70.9% `MaxBoxAcc` and `PxAP`). We attribute the high FSL performance on CUB to the fact that all images are birds; with 1 sample/class, there are effectively 200 birds as training samples. For OpenImages, the high FSL performance is due to the rich supervision provided by pixel-wise masks. On ImageNet, FSL results are not as great: they surpass the CAM result at 8-10 samples per class. Overall, however, FSL performances are strikingly good, even at a low data regime. Thus, given a few fully labeled samples, it is perhaps better to train a model with it than to search hyperparameters. Only when there is absolutely no full supervision (0 fully labeled sample), CAM is meaningful (better than the no-learning center-gaussian baseline).

6. Discussion and Conclusion

After years of weakly-supervised object localization (WSOL) research, we look back on the common practice and make a critical appraisal. Based on a precise definition of the task, we have argued that WSOL is ill-posed and have discussed how previous methods have used different types of implicit full supervision (*e.g.* tuning hyperparameters with pixel-level annotations) to bypass this issue (§3). We have then proposed an improved evaluation protocol that allows the hyperparameter search over a few labeled samples (§4). Our empirical studies lead to some striking conclusions: CAM is still not worse than the follow-up methods (§5.2) and it is perhaps better to use the full supervision directly for model fitting, rather than for hyperparameter search (§5.5).

We propose the following future research directions for the field. (1) Resolve the ill-posedness via *e.g.* adding more background-class images (§3.2). (2) Define the new task, *semi-weakly-supervised object localization*, where methods incorporating both weak and full supervision are studied.

Our work has implications in other tasks where learners are not supposed to be given full supervision, but are supervised implicitly via model selection and hyperparameter fitting. Examples include weakly-supervised vision tasks (*e.g.* detection and segmentation), zero-shot learning, and unsupervised tasks (*e.g.* disentanglement [30]).

Acknowledgements. The work is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (NRF-2019R1A2C2006123) and ICT R&D program of MSIP/IITP [R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding]. This work was also funded by DFG-EXC-Nummer 2064/1-Projektnummer 390727645 and the ERC under the Horizon 2020 program (grant agreement No. 853489).

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 4

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, pages 549–565, 2016. 2

[3] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, pages 11700–11709, 2019. 5

[4] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012. 5

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 5

[6] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, pages 2219–2228, 2019. 1, 2, 3, 4, 5, 6

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 2

[8] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, pages 1635–1643, 2015. 2

[9] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *ECCV*, pages 452–466. Springer, 2010. 2

[10] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100(3):275–293, 2012. 4

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2

[12] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pages 3429–3437, 2017. 2

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 2

[14] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, pages 2409–2416, 2014. 2, 3

[15] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *ICML*, 2019. 2

[16] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2019. 2

[17] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017. 2

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[20] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *ECCV*, 2018. 2

[21] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NIPS*, pages 1495–1503, 2015. 2

[22] James D Keeler, David E Rumelhart, and Wee Kheng Leow. Integrated segmentation and recognition of hand-printed numerals. In *NIPS*, pages 557–563, 1991. 3

[23] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, pages 876–885, 2017. 5

[24] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self driving vehicles. In *ECCV*, 2018. 2

[25] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, pages 3524–3533, 2017. 1, 2, 3, 4, 5, 6

[26] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, pages 3159–3167, 2016. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2

[28] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1307, 2019. 2

[29] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2010. 5

[30] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, pages 4114–4124, 2019. 8

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2, 5

[32] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency

for object segmentation from image level labels. In *CVPR*, pages 5038–5047, 2017. 5

[33] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, pages 685–694, 2015. 2

[34] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *ECCV*, pages 361–376, 2014. 2

[35] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, pages 1742–1750, 2015. 2, 3

[36] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, 2018. 2

[37] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 2

[38] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019. 5

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *SIGKDD*, pages 1135–1144, 2016. 2

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, 2018. 2

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 2, 3, 4

[42] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 2

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 2

[44] Miaojing Shi and Vittorio Ferrari. Weakly supervised object localization using size estimates. In *ECCV*, pages 105–121. Springer, 2016. 2, 3

[45] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR (workshop track)*, 2014. 2

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[47] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *ICML*, pages 1611–1619, 2014. 2, 3

[48] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 2

[49] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328, 2017. 2

[50] AH Sung. Ranking importance of input parameters of neural networks. *Expert Systems with Applications*, 15(3-4):405–411, 1998. 2

[51] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 5

[52] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1, 5

[53] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. In *CVPR*, 2019. 2

[54] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *ICCV*, pages 6589–6598, 2019. 2

[55] 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 *ICCV*, 2019. 1, 2, 3, 4, 5, 6

[56] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, pages 1325–1334, 2018. 1, 2, 3, 4, 5, 6

[57] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *ECCV*, pages 597–613, 2018. 1, 2, 3, 4, 5, 6

[58] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 1, 2, 3, 4, 5, 6

[59] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *ICLR*, 2017. 2