On Label Granularity and Object Localization

Elijah Cole[1,†] Kimberly Wilber[2] Grant Van Horn[3] Xuan Yang[2]
Marco Fornoni[2] Pietro Perona[1] Serge Belongie[4]
Andrew Howard[2] Oisin Mac Aodha[5]

[1]Caltech [2]Google [3]Cornell University
[4]University of Copenhagen [5]University of Edinburgh

[†]ecole@caltech.edu

Abstract. Weakly supervised object localization (WSOL) aims to learn representations that encode object location using only image-level category labels. However, many objects can be labeled at different levels of granularity. Is it an animal, a bird, or a great horned owl? Which image-level labels should we use? In this paper we study the role of label granularity in WSOL. To facilitate this investigation we introduce iNatLoc500, a new large-scale fine-grained benchmark dataset for WSOL. Surprisingly, we find that choosing the right training label granularity provides a much larger performance boost than choosing the best WSOL algorithm. We also show that changing the label granularity can significantly improve data efficiency.

1 Introduction

For many problems in computer vision, it is not enough to know *what* is in an image, we also need to know *where* it is. Examples can be found in many domains, including ecological conservation [20], autonomous driving [55], and medical image analysis [30]. The most popular paradigm for locating objects in images is object *detection*, which aims to predict a bounding box for every instance of every category of interest. Object *localization* is special case of detection where each image is assumed to contain exactly one object instance of interest, and the category of that object is known.

Standard approaches to object detection and localization require bounding boxes for training, which are expensive to collect at scale [37]. Weakly supervised object localization (WSOL) methods aim to sidestep this obstacle by learning to localize objects using only image-level labels at training time. The potential reduction in annotation cost which could result from effective weakly supervised methods has stimulated significant interest in WSOL over the last few years [60].

In this paper we explore the role of label granularity in WSOL. The *granularity* of a category is the degree to which it is specific, which can vary from coarse-grained (e.g. "animal") to fine-grained (e.g. "great horned owl") [54]. When we work with benchmark datasets in computer vision, we often take the given level of label granularity for granted. However, it is usually possible to make those

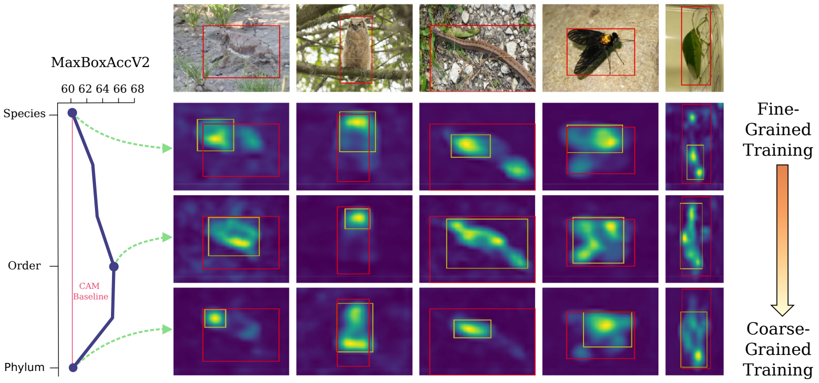

Fig. 1. Label granularity is a critical but understudied factor in weakly supervised object localization (WSOL). We show five hand-picked examples from our iNatLoc500 dataset. Below each image we show class activation maps (CAMs) [63] derived from training a classifier at different granularity levels, with ground truth bounding boxes (red) and WSOL-based bounding boxes (yellow) superimposed. Conventional training does not consider label granularity and can lead to inferior localization performance (red line). Better WSOL results can be achieved by training with coarse (i.e. "order") labels, as opposed to fine-grained (i.e. "species") ones.

labels more general or more specific. It is worth asking whether the label granularity we are given is the best one to use for a certain task. Label granularity matters for WSOL because the first step in most WSOL algorithms is to train a classifier using image-level category labels. By choosing a label granularity we are choosing which training images are grouped into categories. This affects the discriminative features learned by the classifier and ultimately determines the bounding box predictions. Is it possible to improve WSOL performance by controlling label granularity?

Unfortunately, it is difficult to explore label granularity in WSOL due to the limitations of existing datasets. The field of WSOL largely relies on CUB [52] and ImageNet [41]. CUB has a consistent label hierarchy (i.e. one that can be used to measure label granularity), but it is small (∼6k training images) and homogeneous (only bird categories). ImageNet is large and diverse, but lacks a consistent label hierarchy (see Sec. 4.2). Furthermore, [12] recently found that many purported algorithmic advances in WSOL over the last few years – which were based on these two datasets – perform no better than baselines when they are evaluated fairly. This calls for the development of more diverse and challenging benchmarks for WSOL.
Our primary contributions are as follows:

1. We explore the effect of label granularity on WSOL, and show that training at coarser levels of granularity leads to surprisingly large performance gains across many different WSOL methods compared to conventional training e.g. +5.1 `MaxBoxAccV2` for CAM and +6.6 `MaxBoxAccV2` for CutMix (see Fig. 3)

2. We demonstrate that training on coarse labels is more data efficient than conventional training. For instance, training at a coarser level achieves the same performance as conventional CAM with $\sim 15\times$ fewer labels (see Fig. 4).
3. We introduce the iNaturalist Localization 500 (iNatLoc500) dataset, which consists of 138k images for weakly supervised training and 25k images with manually verified bounding boxes for validation and testing. iNatLoc500 covers 500 diverse categories with a consistent hierarchical label space.

2 Related Work

Here we primarily focus on literature related to WSOL. See [60] for a broader overview of related techniques such as weakly supervised object detection [6,7,47].
Weakly Supervised Object Localization. The goal of WSOL is to determine the location of single objects in images using only image-level labels at training time. Early attempts at WSOL explored a variety of different approaches, such as adapting boosting-based methods [34], framing the problem as multiple instance learning [19,21], and applying latent deformable part-based formulations [36].

Some foundational work in deep learning investigated the degree to which object localization comes "for free" when training supervised CNNs for image classification tasks [59,35,63]. In particular, the Class Activation Mapping (CAM) method of [63] showed that CNNs can capture some object location information even when they are trained using only image-level class labels. This inspired a large body of work (e.g. [61,45,62,13,26,27]) that attempted to address some of the shortcomings of CAM, e.g. by preventing the underlying model from only focusing on the most discriminative parts of an object [58] or increasing the spatial resolution of its outputs [43,10].

Recently, [12] showed that when state-of-the-art WSOL methods are fairly compared (e.g. by controlling for the backbone architecture and operating thresholds), they are no better than the standard CAM [63] baseline. Thus, despite its simplicity, CAM is still a surprisingly effective baseline for WSOL. Subsequent work has explored further techniques for improving CAM-based methods [2,28] and alternative approaches for estimating model coefficients [24].
Task Granularity and Localization. Despite the considerable interest in WSOL in recent years, many open questions remain. Examples include the effect of label granularity (e.g. coarse-grained labels like "bird" vs. fine-grained labels indicating the specific species of bird) and the effect of training set size. In the context of supervised object detection, [51] showed that *coarsening* category labels at training time can improve the localization performance of *object detectors*. It is unclear if the same phenomenon holds for WSOL. [53] explored the impact of label granularity for object detection on the OpenImages [29] dataset and observed a small performance improvement when training on finer labels. In the semi-supervised detection setting, [57] trained object detectors on OpenImages and ImageNet using both coarse-grained bounding box annotations and fine-grained image-level labels. [49] also explored semi-supervised detection with an approach that generates object proposals across multiple hierarchical

levels. Unlike our work, these detection-based methods require bounding box information at training time. In addition, the label hierarchies for datasets like ImageNet and OpenImages are not necessarily good proxies for visual similarity or concept granularity (see Sec. 4.2).

For WSOL, [27] showed that aggregating class attribution maps at coarser hierarchical levels (e.g. “dog”) results in more spatial coverage of the objects of interest, whereas maps for finer-scale concepts (e.g. “Afghan hound”) only focus on subparts of the object. However, their analysis does not explore the impact of training at different granularity levels. It is also worth noting that their aggregation method only improves performance on CUB. Regarding data quantity, [12] studied the number of supervised examples used to tune the hyperparameters of CAM, but did not consider the impact of the number of examples used to train the image classifier.

Though not directly related to our work, we note that label granularity has been studied in many contexts other than object localization, including action recognition [44], knowledge tracing [14], animal face alignment [25], and fashion attribute recognition [22]. In the context of image classification, prior work has tackled topics like analyzing the emergence of hierarchical structure in trained classifiers [5], identifying patterns in visual concept generalization [42], and training finer-grained image classifiers using only coarse-grained labels [46,40,56,48].

Datasets for Object Localization. Early work in WSOL (e.g. [34,19,32]) focused on relatively simple and small-scale datasets such as Caltech4 [18], the Weizmann Horse Database [8], or subsets of PASCAL-VOC [17]. With the rise of deep learning-based methods, CUB [52] and ImageNet [16,41] became the standard benchmarks for this task. CUB [52] consists of images of 200 different categories of birds, where each image contains a single bird instance. ImageNet [16,41] contains 1000 diverse categories and has significantly more images than CUB (>1M compared to ∼6k). [12] proposed OpenImages30k, a 100-category localization-focused subset of the OpenImages V5 dataset [29]. An overview of these datasets is presented in Table 1.

These existing datasets are valuable, but they have shortcomings. CUB is small and homogeneous (only birds). OpenImages30k, as presented in [12], is not actually evaluated as a bounding box localization task. It is instead a per-pixel foreground object segmentation task where the ground truth also features some “ignore” regions that are excluded from the evaluation. Finally, while both OpenImages30k and ImageNet have label hierarchies, they do not reflect concept granularity in a consistent way. As a result, it is difficult to use them to better understand the relationship between concept granularity and localization. We discuss these issues in greater detail in Sec. 4.2. To address these shortcomings we introduce iNatLoc500, a new WSOL dataset composed of images from 500 fine-grained visual categories and equipped with a consistent label hierarchy.

Table 1. Comparison of datasets for WSOL. The vast majority of WSOL papers use only CUB and ImageNet. The OpenImages30k dataset was introduced by [12], which also defines the splits we use for CUB and ImageNet. For each split we provide the minimum, maximum, and mean number of images per category, along with the total number of images in the split. Means are rounded to the nearest integer. The properties of these four datasets are discussed in detail in Sec. 4.2.

		`train-weaksup` (D_w)				`train-fullsup` (D_f)				`test` (D_{test})			
Dataset	# Cat.	Min	Max	Mean	Total	Min	Max	Mean	Total	Min	Max	Mean	Total
CUB [52]	200	29	30	30	6k	3	6	5	1k	11	30	29	5.8k
ImageNet [16]	1000	732	1300	1281	1.28M	10	10	10	10k	10	10	10	10k
OpenImages30k [3,12]	100	230	300	298	30k	25	25	25	2.5k	50	50	50	5k
iNatLoc500	500	149	307	276	138k	25	25	25	12.5k	25	25	25	12.5k

3 Background

3.1 Weakly Supervised Object Localization (WSOL)

We begin by formalizing the WSOL setting. Let D_w be a set of *weakly labeled* images, i.e. $D_w = \{(x_i, y_i)\}_{i=1}^{N_w}$ where $x_i \in \mathbb{R}^{H \times W \times 3}$ is an image and $y_i \in \{1, \ldots, C\}$ is an image-level label corresponding to one of C categories. Let D_f be a set of *fully labeled* images, i.e. $D_f = \{(x_i, y_i, \mathbf{b}_i)\}_{i=1}^{N_f}$ where x_i and y_i are defined as before and $\mathbf{b}_i \in \mathbb{R}^4$ is a bounding box for an instance of category y_i. In practice $N_w \gg N_f$. WSOL approaches typically comprise three steps:
(1) Train. Use D_w to train an image classifier $h_\theta : \mathbb{R}^{H \times W \times 3} \rightarrow [0,1]^C$ by solving

$$\hat{\theta}(D_w) = \text{argmin}_\theta \frac{1}{|D_w|} \sum_{(x_i, y_i) \in D_w} \mathcal{L}(h_\theta(x_i), y_i)$$

where $\mathcal{L}$ is some training loss and θ represents the parameters of h. Different WSOL methods are primarily distinguished by the loss functions and training protocols they use to train h.
(2) Localize. For each $(x_i, y_i, \mathbf{b}_i) \in D_f$, predict a bounding box

$$\hat{\mathbf{b}}_i = g(x_i, y_i | h_{\hat{\theta}(D_w)})$$

according to some procedure $g : \mathbb{R}^{H \times W \times 3} \times \{1, \ldots, C\} \rightarrow \mathbb{R}^4$. Typically g is a simple sequence of image processing operations applied to the feature maps of the trained classifier $h_{\hat{\theta}(D_w)}$.
(3) Evaluate. Let E denote a suitable WSOL error metric which compares the predicted boxes $\{\hat{\mathbf{b}}_i\}_{i=1}^{N_f}$ against the ground-truth boxes $\{\mathbf{b}_i\}_{i=1}^{N_f}$. Use the validation error $E(D_f | D_w)$ for model selection and hyperparameter tuning and then use a held-out test set D_{test} (which is fully labeled like D_f) to measure test error $E(D_{\text{test}} | D_w)$. See [12] for a discussion of WSOL performance metrics.

The role of low-shot supervised localization. Without the fully labeled images D_f, the WSOL problem becomes ill-posed [12]. Since WSOL therefore requires at least a small number of bounding box annotations for validation, it is natural to ask how WSOL compares to few-shot object localization? For our purposes, we define few-shot object localization methods as those which use only D_f for training and validation. Under this definition, the few-shot methods (which use only D_f) actually require strictly *less* data than WSOL (which requires both D_w and D_f). Since WSOL and few-shot object localization are practical alternatives, it is important to consider them together as in [12].

3.2 Label Hierarchies and Label Granularity

We define a *label hierarchy* (on a label set L) to be a directed rooted tree H whose leaf nodes (i.e. nodes $v \in H$ with no children) correspond to the labels in L. Edges in H represent "is-a" relationships, so a directed edge from $u \in H$ to $v \in H$ means that v (e.g. "bird") is a kind of u (e.g. "animal"). We overload L to refer to the label set and to the corresponding set of nodes in H. Let r denote the root node of H and let $d(u, v)$ denote the number of edges on the path from $u \in H$ to $v \in H$.

Coarsening a label. Because there is a unique path from the root node r to any leaf node $\ell \in L$, we can "coarsen" the label ℓ in a well-defined way by merging it with its parent node. We define the *coarsening operator* $c_k : H \to H$, which takes any node in the label hierarchy and returns the node which is k edges closer to the root. Thus, $c_0(\ell) = \ell$, $c_1(\ell)$ is the parent of ℓ, $c_2(\ell)$ is the grandparent of ℓ, and so on, with $c_k(\ell) = r$ for all $k \geq d(r, \ell)$.

Coarsening a dataset. We can describe a general "coarsened" version of $D_w = \{(x_i, y_i)\}_{i=1}^{N_w}$ as $D_w^{\mathbf{k}} = \{(x_i, c_{k_i}(y_i))\}_{i=1}^{N_w}$ where $\mathbf{k} = (k_1, \ldots, k_{|D_w|})$. If we allow the entries of $\mathbf{k}$ to be chosen completely independently, then we can encounter problems e.g. images with multiple valid labels. To prevent these cases, we require $\mathbf{k}$ to be chosen such that $c_{k_i}(y_i) \in H$ is not a descendant of $c_{k_j}(y_j) \in H$ for any $i, j \in \{1, \ldots, N_w\}$.

Problem statement. We can now formalize our key questions: How does $\mathbf{k}$ affect $E(D_{\text{test}} | D_w^{\mathbf{k}})$? Are there choices of $\mathbf{k}$ such that $E(D_{\text{test}} | D_w^{\mathbf{k}}) < E(D_{\text{test}} | D_w)$?

4 The iNatLoc500 Dataset

In this section we introduce the iNaturalist Localization 500 (iNatLoc500) dataset, a large-scale fine-grained dataset for weakly supervised object localization. We first detail the process of building the dataset and cleaning the localization annotations. We then discuss the key properties of the dataset and highlight the advantages of iNatLoc500 compared to three WSOL datasets that are currently commonly used (CUB, ImageNet, and OpenImages30k).

iNatLoc500 has three parts: `train-weaksup` (D_w), `train-fullsup` (D_f), and `test` (D_{test}). Each image in the weakly supervised training set (D_w) has one image-level category label. Each image in the fully supervised validation set (D_f) and test set (D_{test}) has one image-level category label *and* one bounding box annotation. All bounding boxes have been manually validated. Split statistics

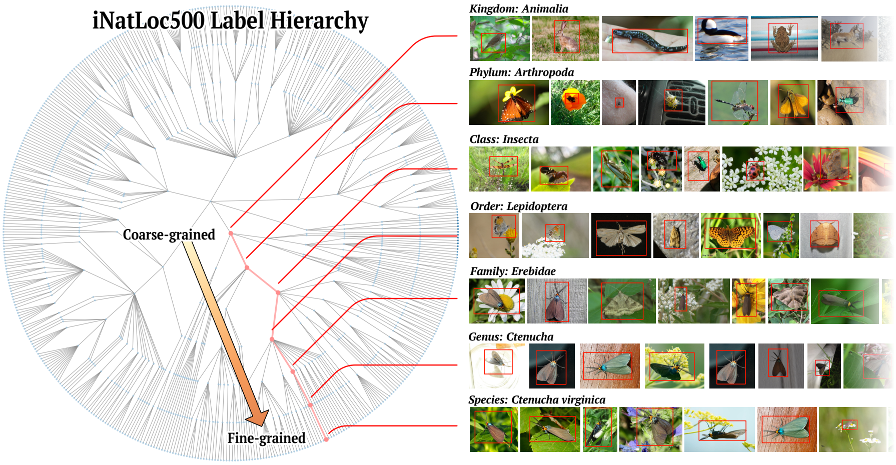

Fig. 2. Sample images from iNatLoc500 at different levels of the label hierarchy, from coarse ("kingdom") to fine ("species"). Random images from coarse levels of the hierarchy tend to be much more varied than random images ones from finer levels.

are presented in Table 1 and sample images from the dataset can be found in Fig. 2. The dataset is publicly available.[1]

4.1 Dataset Construction

The iNatLoc500 dataset is derived from two existing datasets: iNat17 [51] and iNat21 [50]. Both datasets contain images of plants and animals collected by the citizen science platform iNaturalist [1]. iNat21 is much larger than iNat17 (2.7M images, 10k species vs. 675k images, 5k species), but iNat17 has crowdsourced bounding box annotations. We draw from iNat21 for D_w and we draw from iNat17 for D_f and D_{test}.

Full details on the process of constructing iNatLoc500 can be found in the supplementary material, but we note two important design choices here. First, iNat17 did not collect bounding boxes for plant categories because it is often unclear how to draw bounding boxes for plants. Consequently, iNatLoc500 does not contain any plant categories. Second, we set very high quality standards for the bounding boxes. Five computer vision researchers manually reviewed $\sim$ 65k images to ensure the quality of the bounding boxes for D_f and D_{test}, of which only 51% met our quality standards. Explicit quality criteria and examples of removed images can be found in the supplementary material.

4.2 Dataset Properties

iNatLoc500 is fine-grained, large-scale, and visually diverse. Moreover, iNatLoc500 has a consistent label hierarchy which serves as a reliable proxy for

[1] https://github.com/visipedia/inat_loc/

label granularity. We now discuss the importance of each of these properties and contrast iNatLoc500 with existing WSOL datasets.

Fine-grained categories. Each category in iNatLoc500 corresponds to a different species, and the differences between species can be so subtle as to require expert-level knowledge [51]. While there are challenging images in ImageNet and OpenImages30k, most of the categories are coarse-grained i.e. relatively few pairs of categories are highly visually similar. For instance, the reptile categories in OpenImages30k (`lizard`, `snake`, `frog`, `crocodile`) are typically easy to distinguish. In iNatLoc-500 there are 107 reptile species, some of which are highly similar (e.g. `Chihuahuan spotted whiptail` vs. `Common spotted whiptail`).

Consistent label hierarchy. The label hierarchy for iNatLoc500 consists of the following seven tiers, ordered from coarsest to finest: kingdom, phylum, class, order, family, genus, and species. All of the species in iNatLoc500 are animals, so the "kingdom" tier only has one node (`Animalia`), which is the root node of the label hierarchy. Every species lies at the same distance from the root. The iNatLoc500 label hierarchy is *consistent* in the sense that all nodes at a given level of the hierarchy correspond to concepts with similar levels of specificity. This means that depth in the label hierarchy measures label granularity. The label hierarchy for CUB is also consistent. However, the taxonomies that underlie ImageNet and OpenImages30k are considerably more arbitrary. For instance, in OpenImages30k some categories are far from the root of the label hierarchy (e.g. `entity/vehicle/land_vehicle/car/limousine` or `entity/animal/mammal/carnivore/fox`) while others are close to the root (e.g. `entity/bicycle_wheel` or `entity/human_ear`) despite the fact that there is no obvious difference in concept specificity.

Unambiguous label semantics. The categories in iNatLoc500 are well-defined in the sense that (for most species) there is little room for debate about what "counts" as an instance of that species. While the distinctions between species can be quite subtle, each species is a well-defined category. CUB shares this advantage for the most part, but ImageNet and OpenImages30k do not. For instance, OpenImages30k contains the categories `wine` and `bottle`. To which category does a bottle of wine belong? (In fact, we find bottles of wine in both categories.) ImageNet is known to have similar issues with ambiguous and overlapping category definitions [4].

Visual diversity. Like ImageNet and OpenImages30k, iNatLoc500 has a category set which exhibits a high degree of visual diversity. CUB is much more homogeneous, consisting of only birds. Combined with its consistent label hierarchy, the visual diversity of iNatLoc500 enables future work on e.g. how localization ability generalizes across categories as a function of taxonomic distance.

Large scale. iNatLoc500 is a large-scale dataset, both in terms of the number of categories and the number of training images. CUB and OpenImages30k are considerably smaller on both counts. Large training sets are valuable because they simplify supervised learning. Large training sets also enable research on self-supervised representation learning, which has received little attention thus far in WSOL. We provide a summary of the key dataset statistics in Table 1.

5 Experiments

In this section we present WSOL results on iNatLoc500 as well as existing benchmark datasets. We also consider few-shot learning baselines based on segmentation and detection architectures. Finally, we use the unique properties of iNatLoc500 to study how label granularity affects localization performance and data efficiency. A summary of the different WSOL datasets can be found in Table 1.

5.1 Implementation Details

Performance metrics. All WSOL performance numbers in this paper are `MaxBoxAccV2`, which is defined in [12]. The only exceptions are the results for OpenImages30k in Table 2, which are given in `PxAP` as defined in [12].
Fixed-granularity training. In Sec. 5.3 we probe the effect of granularity on WSOL by training on "coarsened" versions of D_w. In the notation of Sec. 3.2, these can be written $D_w^{k\cdot\mathbf{1}}$ for $k = 1, 2, \ldots$, where $\mathbf{1}$ denotes the "all ones" vector. This corresponds to merging all leaves with their parent k times. We then run the entire WSOL pipeline from scratch to compute $E(D_{\text{test}}|D_w^{k\cdot\mathbf{1}})$ for each k. To the best of our knowledge this is compatible with all existing WSOL methods.
Fixed-granularity CAM aggregation. We also consider a second method for using label hierarchy information to improve WSOL, inspired by [27]. Just like traditional CAM, the first step is to train an image classifier using the standard (most fine-grained) label set. However, instead of returning only the CAM for the species labeled in the input image, we return a CAM for each species in the same genus / family / ... / phylum and average them. This "aggregated" CAM is then evaluated as normal. We abbreviate this method as CAM-Agg.
Hyperparameter search for WSOL methods. Each time we train a WSOL method we re-tune the learning rate over the set $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and choose the one that leads to the best `MaxBoxAccV2` performance on the fully supervised validation set D_f. We then report the `MaxBoxAccV2` performance for the selected model on D_{test}. We leave all other hyperparameters fixed. Full training details can be found in the supplementary material.
Non-WSOL methods. We provide results for the baselines proposed in [12] (Center, FSL-Seg), as well as a new few-shot detection baseline (FSL-Det). "Center" is a naive baseline that simply assumes a centered Gaussian activation map for all images. "FSL-Seg" is a supervised baseline that is trained on the D_f split of each dataset. The architecture is based on models for saliency mask prediction [33]. Finally, we introduce "FSL-Det", a few-shot detection baseline for WSOL that is also trained on D_f. It uses Faster-RCNN [39] with the same backbone as other methods (i.e. ImageNet-pretrained ResNet-50 [23]). Full implementation details can be found in the supplementary material.

5.2 Baseline Results

We follow [12] and evaluate six recent WSOL methods and two non-WSOL methods (Center and FSL-Seg) on iNatLoc500. The results can be found in Table 2.

Table 2. Comparison of WSOL methods. Numbers are `MaxBoxAccV2` for ImageNet, CUB, and iNatLoc500 and `PxAP` for OpenImages30k. All results use an ImageNet-pretrained ResNet-50 [23] backbone with an input resolution of 224x224. WSOL numbers for ImageNet, CUB, and OpenImages30k are the updated results from [11]. WSOL numbers for iNatLoc500 are our own, as are the numbers for the baselines (Center, FSL-Seg, FSL-Det). FSL baselines use 10 images / class for ImageNet, 5 images / class for CUB, 25 images / class for OpenImages30k, and 25 images / class for iNatLoc500. We do not report FSL-Det for OpenImages30k because the evaluation protocol for that dataset requires segmentation masks.

Method	ImageNet	CUB	OpenImages30k	iNatLoc500
CAM [63]	63.7	63.0	58.5	60.2
HaS [45]	63.4	64.7	55.9	60.0
ACoL [61]	62.3	66.5	57.3	55.3
SPG [62]	63.3	60.4	56.7	60.7
ADL [13]	63.7	58.4	55.2	58.9
CutMix [58]	63.3	62.8	57.7	60.1
Center	53.4	56.8	46.0	42.8
FSL-Seg	68.7	89.4	75.2	78.6
FSL-Det	70.4	95.4	-	83.6

We focus our observations on ImageNet, CUB, and iNatLoc500 since OpenImages30k is evaluated using a different task and evaluation metric. We first note that our findings on iNatLoc500 reinforce the main results from [12], namely that (a) none of the WSOL methods performs substantially better than CAM and (b) FSL-Seg significantly outperforms all WSOL methods. Second, if we consider the performance gap between CAM and the Center baseline, we see that simple centered boxes are not as successful on iNatLoc500 (-17.2 `MaxBoxAccV2`) as they are on CUB (-6.2 `MaxBoxAccV2`) and ImageNet (-10.3 `MaxBoxAccV2`). This indicates that iNatLoc500 is a more challenging dataset for benchmarking WSOL. Finally, we provide results for our few-shot detection baseline (FSL-Det). For ImageNet, CUB, and iNatLoc500 we find that FSL-Det is a stronger baseline than FSL-Seg. Like FSL-Seg, FSL-Det directly trains on the boxes in D_f, whereas the WSOL methods only use those boxes to tune their hyperparameters. However, FSL-Det sets a new ceiling for localization performance on these datasets, indicating that current WSOL methods have considerable room for improvement.

5.3 Label Granularity and Localization Performance

iNatLoc500 is equipped with a consistent label hierarchy which allows us to directly study the relationship between label granularity and localization performance. The traditional approach to WSOL on iNatLoc500 would begin by training a classifier on the *species-level* labels, i.e. the finest level in the label hierarchy. However, our hypothesis is that training at the most fine-grained level may not lead to the best localization performance. To study this, we use the fixed-granularity training method discussed in Sec. 5.1. In particular, we “re-label” D_w at each level of the label hierarchy using successively coarser categories. We then use each of these re-labeled datasets to train and evaluate different WSOL methods. The results in Fig. 3(left) show that coarsening the labels of D_w can significantly boost WSOL performance (e.g. up to +5.1 `MaxBoxAccV2` for CAM). The numerical values plotted in Fig. 3(left) can be found in the supplementary materials. Note that it would be difficult to draw similar conclusions by studying ImageNet or OpenImages30k because their label hierarchies do not measure how fine-grained different categories are – see Sec. 4.2 for a discussion. Our conceptually simple coarsening approach results in large performance improvements across five different WSOL methods, without any modifications to the model architectures or training losses.

Coarse training beyond iNatLoc500. Fig. 3(left) shows that coarse training significantly improves WSOL performance on iNatLoc500. We study the effect of coarse training on FGVC-Aircraft [31], CUB [52], and ImageNet [16] in the supplementary material. As expected, FGVC-Aircraft and CUB (which have consistent label hierarchies) both benefit from coarse training while ImageNet (which lacks a consistent label hierarchy) does not.

Localization performance vs. classification performance. In Fig. 3(right) we show the image classification performance for each WSOL method in Fig. 3(left) at each granularity level. We see that classification performance and WSOL performance are not necessarily correlated. WSOL performance increases before decreasing at the coarsest level of granularity. Classification performance increases with label coarsening, even at the coarsest level of granularity.

An alternative method for incorporating label granularity. We also present the performance of CAM-Agg, an alternative method for incorporating granularity information in WSOL (see Sec. 5.1). In our experiments, CAM-Agg underperforms vanilla CAM at every granularity level. As a point of comparison, [27] finds that CAM-Agg is better than CAM for CUB but worse than CAM for ImageNet. Our findings suggest that training the model with coarse categories leads to much better localization performance when compared to aggregating the localization outputs for multiple similar fine-grained categories.

5.4 Label Granularity and Data Efficiency

Most WSOL work makes D_w as large as possible by default, so there has been little attention paid to how the size of D_w trades off against localization performance. In this section we analyze the performance of CAM-based WSOL as a function of the size of D_w. We are particularly interested in how label granularity interacts with data efficiency. To study this question, we first pick a granularity level and generate subsampled versions of D_w by choosing, uniformly at random, 50, 100, or 200 images from each category. Note that the size of each subsampled version of D_w depends on the granularity level. For instance, if the categories are the 317 genera, then 50 images per category is $50 \times 317 = 15,850$, compared to $50 \times 61 = 3,050$ images if the categories are the 61 orders. We present WSOL results for four granularity levels in Fig. 4. We find that by training at a coarser level, we can obtain better performance with fewer labels. All of the square markers above the dashed line in Fig. 4 correspond to cases where we can

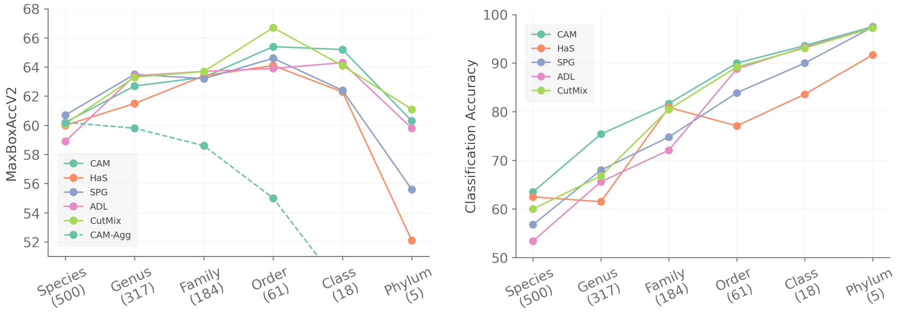

Fig. 3. Effect of label granularity of D_w on WSOL performance (left) and classification accuracy (right) for iNatLoc500. The number of categories at each tier is given in parentheses. **(Left)** Localization performance suffers when the category labels are either too fine (e.g. Species) or too coarse (e.g. Phylum). The results on the very left of the plot are the same as those in Table 2. Note that ACoL is excluded due to poor performance – we suspect it requires more epochs of training than the standard protocol allows for iNatLoc500. We also show results for CAM-Agg (Sec. 5.1), an alternative method for aggregating hierarchy information in WSOL. **(Right)** Each WSOL method trains the image classifier in a different way, but classification accuracy generally increases as the labels become more coarse. Naturally it is easier to distinguish between coarser categories, but it is interesting to note that classification performance is excellent at the phylum level, despite poor localization performance.

achieve better performance than the standard species-level CAM approach using fewer labels. To take one example, by training at the family level we can match the performance of the standard CAM approach by training with 50 images per family (9200 images), a training set reduction of ∼15×.

6 Discussion

Why does performance increase as we coarsen the labels? In Fig. 3(left) we see that five different WSOL algorithms perform better as we coarsen the labels in D_w, up until the coarsest level when performance drops. What accounts for this behavior? Our analysis of CAM in Fig. 5 provides some clues. Fig. 5(left) shows that the area of the predicted box tends to be larger than the area of the ground truth box, and that their ratio *decreases* towards unity as we coarsen the labels (black curve). That is, the predicted box size gets closer to the true box size as we coarsen the labels. This casts doubt on a common intuition (which as far as we know has not been empirically investigated before now) that WSOL methods predict smaller boxes for more fine-grained categories [27].

Why does performance drop at the coarsest level of granularity? In Fig. 5(left) we see that as we coarsen the labels the concentration of *activation* in the ground truth box *increases* before collapsing at the coarsest level (red curve). Fig. 5(right) shows that the activation maps become highly fragmented

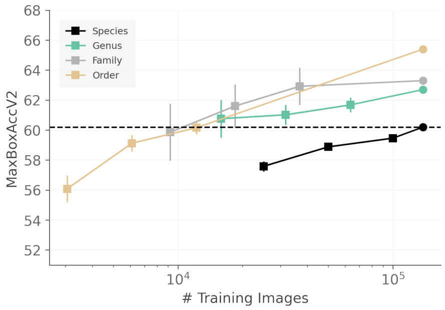

Fig. 4. Effect of the number of training images (N_w) on CAM performance for iNatLoc500. The dashed line corresponds to the performance of species-level CAM with the entirety of D_w. Each color corresponds to a different label granularity for D_w. Circles at the right of the graph indicate performance using all of D_w. Squares represent subsampled datasets which use a fixed number of images per category: 50, 100, or 200. All squares have error bars indicating the standard deviation over 5 runs with different randomly sampled subsets of D_w.

at coarser levels. Taken together, these two findings suggest that at the coarsest level the activation maps tend to focus more on global image characteristics (e.g. land vs. water) than the properties of the foreground object. Note that these features are still useful for image classification, as is shown in Fig. 3(right).

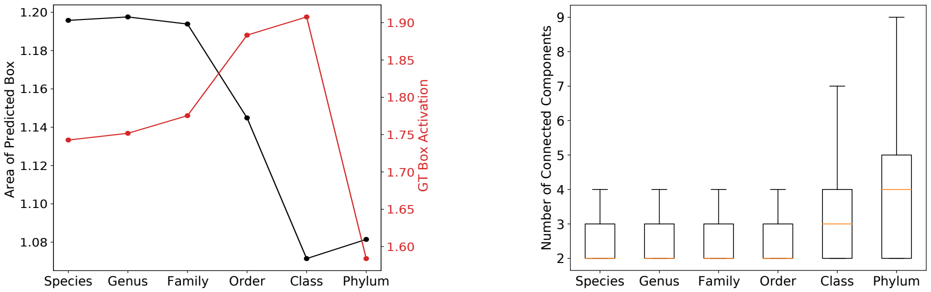

Fig. 5. Analysis of CAM-based WSOL on the D_f split of iNatLoc500. **(Left)** *Black*: Ratio of the area of the predicted box to the area of the ground truth box. *Red*: Ratio of the activation inside the ground truth box to the activation of background pixels. Both curves show medians over the 12.5k images in D_f at each granularity level. **(Right)** Number of connected components in the binarized activation maps at each granularity level. Each box plot shows the distribution over the 12.5k images in D_f. See the supplementary material for full details on the construction of these plots.

Limitations. The iNatLoc500 dataset has several limitations. First, it contains only animal categories. These categories are highly diverse, but they are not representative of all visual domains. Second, it is possible that there are errors in the image-level labels provided by the iNaturalist community, though this is expected to be rare as each image has been labeled by multiple people [50]. Third, many real fine-grained problems have a long-tailed class distribution but, like other localization datasets, iNatLoc500 is approximately balanced (at the species level). Finally, there is a conceptual limitation in our experiments: the use of a single granularity level across the entire dataset. In fact, it is likely that different images are best treated at different granularity levels. Our work does not address this important topic which we leave for future work.

iNatLoc500 can be used to investigate numerous research agendas beyond traditional WSOL. For example, D_w was designed to be large enough for self-supervised learning, which has received surprisingly little attention in the WSOL community [9]. We are also interested in using iNatLoc500 to study whether self-supervised learning methods can be improved by using WSOL methods to select crops [38], especially in the context of fine-grained data [15]. For the object detection community, the clean boxes in iNatLoc500 can (i) serve as a test set for object detectors trained on the noisy iNat17 boxes, (ii) be used to study the problem of learning multi-instance detectors from one box per image, and (iii) be used to analyze the role of label granularity in object detection. Finally, we have seen that hierarchical reasoning can significantly improve localization performance. In the future, we aim to explore methods for automatically determining the most appropriate level of coarseness required for generating representations that best encode object location.

7 Conclusion

We have shown that substantial improvements in WSOL performance can be achieved by modulating the granularity of the training labels, and that coarser-grained training leads to more data-efficient WSOL. We also presented iNatLoc500, a new large-scale fine-grained dataset for WSOL. Despite the gains in performance from coarse-level training, iNatLoc500 remains a challenging localization task which we hope will motivate additional progress in WSOL.

Acknowledgements. We thank the iNaturalist community for sharing images and species annotations. This work was supported by the Caltech Resnick Sustainability Institute, an NSF Graduate Research Fellowship (grant number DGE1745301), and the Pioneer Centre for AI (DNRF grant number P1).

References

1. iNaturalist, `www.inaturalist.org`, accessed Mar 7 2022
2. Bae, W., Noh, J., Kim, G.: Rethinking class activation mapping for weakly supervised object localization. In: ECCV (2020)
3. Benenson, R., Popov, S., Ferrari, V.: Large-scale interactive object segmentation with human annotators. In: CVPR (2019)
4. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv:2006.07159 (2020)
5. Bilal, A., Jourabloo, A., Ye, M., Liu, X., Ren, L.: Do convolutional neural networks learn class hierarchy? IEEE transactions on visualization and computer graphics **24**(1), 152–162 (2017)
6. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: CVPR (2015)

7. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: CVPR (2016)
8. Borenstein, E., Ullman, S.: Class-specific, top-down segmentation. In: ECCV (2002)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
10. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV (2018)
11. Choe, J., Oh, S.J., Chun, S., Akata, Z., Shim, H.: Evaluation for weakly supervised object localization: Protocol, metrics, and datasets. arXiv:2007.04178 (2020)
12. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: CVPR (2020)
13. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: CVPR (2019)
14. Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., Heo, J.: Ednet: A large-scale hierarchical dataset in education. In: International Conference on Artificial Intelligence in Education. pp. 69–73. Springer (2020)
15. Cole, E., Yang, X., Wilber, K., Mac Aodha, O., Belongie, S.: When does contrastive visual representation learning work? In: CVPR (2022)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)
18. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR (2003)
19. Galleguillos, C., Babenko, B., Rabinovich, A., Belongie, S.: Weakly supervised object localization with stable segmentations. In: ECCV (2008)
20. van Gemert, J.C., Verschoor, C.R., Mettes, P., Epema, K., Koh, L.P., Wich, S.: Nature conservation drones for automatic localization and counting of animals. In: ECCV (2014)
21. Gokberk Cinbis, R., Verbeek, J., Schmid, C.: Multi-fold mil training for weakly supervised object localization. In: CVPR (2014)
22. Guo, S., Huang, W., Zhang, X., Srikhanta, P., Cui, Y., Li, Y., Adam, H., Scott, M.R., Belongie, S.: The imaterialist fashion attribute dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
24. Jung, H., Oh, Y.: Towards better explanations of class activation mapping. In: ICCV (2021)
25. Khan, M.H., McDonagh, J., Khan, S., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G.: Animalweb: A large-scale hierarchical dataset of annotated animal faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6939–6948 (2020)
26. Ki, M., Uh, Y., Lee, W., Byun, H.: In-sample contrastive learning and consistent attention for weakly supervised object localization. In: ACCV (2020)
27. Kim, J.M., Choe, J., Akata, Z., Oh, S.J.: Keep calm and improve visual feature attribution. In: ICCV (2021)
28. Kim, J., Choe, J., Yun, S., Kwak, N.: Normalization matters in weakly supervised object localization. In: ICCV (2021)

29. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. IJCV (2020)
30. Li, Z., Dong, M., Wen, S., Hu, X., Zhou, P., Zeng, Z.: Clu-cnns: Object detection for medical images. Neurocomputing (2019)
31. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. Tech. rep. (2013)
32. Nguyen, M.H., Torresani, L., De La Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In: ICCV (2009)
33. Oh, S.J., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B.: Exploiting saliency for object segmentation from image level labels. In: CVPR (2017)
34. Opelt, A., Pinz, A.: Object localization with boosting and weak supervision for generic object recognition. In: Scandinavian Conference on Image Analysis (2005)
35. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: CVPR (2015)
36. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV (2011)
37. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: ICCV (2017)
38. Peng, X., Wang, K., Zhu, Z., Wang, M., You, Y.: Crafting better contrastive views for siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16031–16040 (2022)
39. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS (2015)
40. Robinson, J., Jegelka, S., Sra, S.: Strength from weakness: Fast learning using weak supervision. In: ICML (2020)
41. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
42. Sariyildiz, M.B., Kalantidis, Y., Larlus, D., Alahari, K.: Concept generalization in visual representation learning. In: ICCV (2021)
43. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
44. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2616–2625 (2020)
45. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: ICCV (2017)
46. Taherkhani, F., Kazemi, H., Dabouei, A., Dawson, J., Nasrabadi, N.M.: A weakly supervised fine label classifier enhanced by coarse supervision. In: ICCV (2019)
47. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: CVPR (2017)
48. Touvron, H., Sablayrolles, A., Douze, M., Cord, M., Jégou, H.: Grafit: Learning fine-grained image representations with coarse labels. In: ICCV (2021)
49. Uijlings, J., Popov, S., Ferrari, V.: Revisiting knowledge transfer for training object class detectors. In: CVPR (2018)
50. Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: CVPR (2021)

51. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: CVPR (2018)
52. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
53. Wang, R., Mahajan, D., Ramanathan, V.: What leads to generalization of object proposals? In: ECCV Workshops (2020)
54. Wei, X.S., Song, Y.Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., Yang, J., Belongie, S.: Fine-grained image analysis with deep learning: A survey. PAMI (2021)
55. Wu, B., Iandola, F., Jin, P.H., Keutzer, K.: Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In: CVPR Workshops (2017)
56. Xu, Y., Qian, Q., Li, H., Jin, R., Hu, J.: Weakly supervised representation learning with coarse labels. In: ICCV (2021)
57. Yang, H., Wu, H., Chen, H.: Detecting 11k classes: Large scale object detection without fine-grained bounding boxes. In: ICCV (2019)
58. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)
59. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
60. Zhang, D., Han, J., Cheng, G., Yang, M.H.: Weakly supervised object localization and detection: A survey. PAMI (2021)
61. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: CVPR (2018)
62. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: ECCV (2018)
63. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)