Scaling Novel Object Detection with
Weakly Supervised Detection Transformers

Tyler LaBonte[1,3]   Yale Song[2*]   Xin Wang[1]   Vibhav Vineet[1]   Neel Joshi[1]

[1]Microsoft Research   [2]Meta AI, FAIR   [3]Georgia Institute of Technology

tlabonte@gatech.edu, yalesong@meta.com, {wanxin, vivineet, neel}@microsoft.com

Abstract

A critical object detection task is finetuning an existing model to detect novel objects, but the standard workflow requires bounding box annotations which are time-consuming and expensive to collect. Weakly supervised object detection (WSOD) offers an appealing alternative, where object detectors can be trained using image-level labels. However, the practical application of current WSOD models is limited, as they only operate at small data scales and require multiple rounds of training and refinement. To address this, we propose the Weakly Supervised Detection Transformer, which enables efficient knowledge transfer from a large-scale pretraining dataset to WSOD finetuning on hundreds of novel objects. Additionally, we leverage pretrained knowledge to improve the multiple instance learning (MIL) framework often used in WSOD methods. Our experiments show that our approach outperforms previous state-of-the-art models on large-scale novel object detection datasets, and our scaling study reveals that class quantity is more important than image quantity for WSOD pretraining.

1. Introduction

Object detection is a fundamental task in computer vision where supervised neural networks have demonstrated remarkable performance [55, 53, 44, 6]. A major factor in the success of these approaches is the availability of datasets with fine-grained bounding box and segmentation annotations [20, 43, 38, 27, 59, 39]. However, in comparison to image classification, the annotation process for object detection is considerably more expensive and time-consuming [50]. We consider weakly supervised object detection (WSOD), which aims to learn object detectors using only image-level category labels (*i.e.*, classification labels).

Previous WSOD models [4, 65] often generate object proposals using a low-precision high-recall heuristic [68, 83], then use multiple instance learning (MIL) [16, 46] to recover high-likelihood proposals. With proposal quality established as a major factor in object detection performance [31], a practical direction is to leverage a *source dataset* with bounding box annotations to transfer semantic (class-aware) [67, 5] or class-agnostic [69, 81] knowledge to a *target dataset* of novel objects. These strategies enable the WSOD model to generate more accurate proposals and classifications by exploiting class and object similarity, respectively, between the source and target datasets.

Though the presence of many classes in the source dataset is posited to be essential for effective transfer [69], current WSOD methods are typically designed for and trained on datasets with few classes. A widely-used setting in the literature is COCO-60 [43, 40] (60 classes) to PASCAL VOC [20] (20 classes), with, to our knowledge, the largest source dataset being ILSVRC-179 [58, 81] (179 classes) and the largest target dataset of novel objects being ILSVRC `val1b` (100 classes).[1] This has two major drawbacks which limit the usage of WSOD models in practice. First, knowledge transfer is most effective when objects in the target dataset have visually similar counterparts in the source dataset. In applications, training on few classes may limit the domains for which knowledge transfer is helpful. Second, current WSOD models perform best with multiple rounds of training and refinement [65, 81] or training an additional semantic model on the target dataset [5]. In addition to the extra computation, these methods require a human to identify, *e.g.*, the optimal number of refinements or pseudo ground truth mining steps, which is unscalable if there are hundreds of novel classes and many downstream tasks.

In contrast, what is desired in practice is similar to the pretraining-finetuning framework which has been standard in classification and fully supervised object detection [28, 8] (though not in WSOD). Specifically, we would like to pretrain a single detection model on a large-scale, annotated source dataset with hundreds of classes, then use this model for weakly supervised finetuning on novel objects with image-level labels. Such a framework would empower

---

[1]Uijlings *et al.* [69] transfer from ILSVRC `val1a` (100 classes) to OpenImages [39] (600 classes), but with significant class overlap.
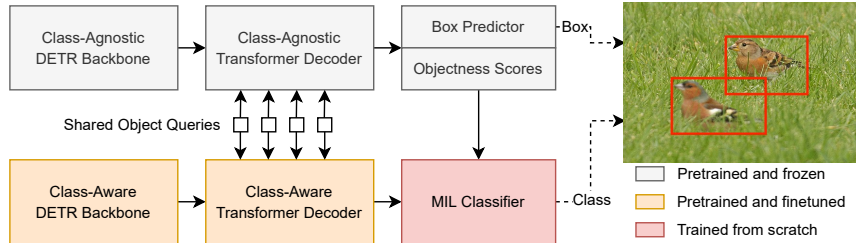
Figure 1: WS-DETR is a hybrid approach utilizing a class-agnostic DETR as proposal generator and a class-aware DETR for weakly supervised finetuning. The two streams share object queries. The MIL classifier leverages objectness knowledge from the pretrained class-agnostic DETR to detect novel objects using only image-level labels.

practitioners to "effortlessly" solve WSOD tasks: the scale and diversity of the source dataset allows for knowledge transfer to a variety of distinct domains, while a simple, single-round finetuning procedure would enable computationally efficient WSOD training without the additional cost of semantic models or multi-round refinement.

To address this practical scenario, we propose the Weakly Supervised Detection Transformer (WS-DETR), which integrates DETR [6] with an MIL architecture for scalable WSOD finetuning on novel objects (detailed in Figure 1). A Transformer-based method is well-suited to this problem because, while they lack the inductive bias of CNNs, they excel at large-scale training and transfer learning for vision tasks [6, 18]. Indeed, the Transformer [70] is the foundation for a widely-used machine learning workflow where massive pretrained models are finetuned to generalize to an incredible variety of downstream tasks, which aligns closely with our desired WSOD framework.

Existing MIL architectures [4] are primarily based on a two-stage Faster R-CNN-like [55] structure where region-of-interest (RoI) pooling is performed on object proposals and the resultant features are passed to a classification head. Our WS-DETR is a hybrid which combines the comprehensive proposal generation of this two-stage framework with the scalability and simplicity of Transformer models.[2] We pretrain a class-agnostic DETR on the source dataset with binary object labels to serve as the proposal generator, then initialize a model with a frozen box predictor from a class-aware DETR pretrained on detection annotations (including classes) for WSOD training. Instead of RoI pooling, we utilize the DETR's learned positional embedding, also called object queries – upon initialization, the object queries are set equal to the class-agnostic object queries and frozen, so the Transformer decoder attends to the same locations as the class-agnostic model and produces classification predictions which correspond to object proposals.

Additionally, we leverage pretrained knowledge to improve MIL training in two ways. First, we show that the

objectness regularization of [81], the state-of-the-art technique for incorporating the object proposal score of the pretrained class-agnostic model, falls prey to the well-known MIL weakness of overfitting to distinctive classification features of novel objects [65]. We propose a new formulation for estimating the joint object and class probabilities which rectifies this error without the need for a box refinement step [65]. Second, since the proposal generator outputs hundreds of object proposals but only a few confident ones, we introduce sparsity along the detection dimension in the MIL classifier, increasing performance by emphasizing correct classification of confident proposals.

We evaluate our WS-DETR's performance on a variety of rigorous and challenging WSOD domains. For large-scale novel object detection, we utilize the Few-Shot Object Detection (FSOD) dataset [21], which has $4\times$ more source classes and $2\times$ more novel target classes than the previous largest datasets used for WSOD [58, 69, 81] and uses a semantic split to maximize novelty. We achieve state-of-the-art performance compared to [81], whose class-agnostic transfer setting is closely related to ours. Our WS-DETR is effective at WSOD training for fine-grained domains – a practical scenario where collecting bounding box labels is difficult [30]. We evaluate fine-grained WSOD performance on the FGVC-Aircraft [45] and iNaturalist [30] datasets and demonstrate state-of-the-art results on up to 2,854 classes. Finally, our scaling study reveals that class quantity is more important than image quantity for WSOD pretraining, which may inform future dataset construction.

2. Related Work

WSOD and MIL. WSOD has been formulated as an MIL problem [16, 46] since well before the deep learning era [22]. Each image is considered as a bag and the object proposals as instances, and a bag is positive for a certain class if it contains at least one instance of that class. The model is provided with candidate proposals, typically via a low-precision high-recall heuristic [68, 83], and learns to identify the most accurate proposals by minimizing the clas-

---

[2]Zhu *et al*. [82] integrate DETR with a two-stage approach in the fully supervised setting; our methods and goals are not directly comparable.

sification error against the bag labels.

The weakly supervised deep detection network (WS-DDN) [4] is a particularly effective method which integrates MIL into an end-to-end learning framework, jointly optimizing a classification layer and a detection layer. Subsequent works have improved WSDDN via self-training and box refinement [35, 15, 65, 80, 77, 66, 75, 79, 37, 56], spatial relations [36, 9, 64], optimization methods [71], residual backbones [61], and segmentation [26, 62, 41]. The MIL classifier in our WS-DETR utilizes a WSDDN-like approach, but we propose using the class-agnostic model's objectness scores to directly compute the joint object and class probabilities rather than learning a detection layer.

Notable WSOD architectures not extending WSDDN include [29, 2, 60, 1]. Self-attention methods have also been proposed [78, 10, 74, 33, 73, 48, 51, 72, 49, 25], though they focus on attention maps rather than Transformer training.

WSOD with Knowledge Transfer. Since initial box quality is critical to WSOD performance [31], using a heuristic to generate proposals is often insufficient and impractical. Several strategies transfer knowledge from fully or partially annotated source datasets whose classes may not coincide with the target objects, including object count annotations [23], semantic relationships [69, 67, 5], segmentation [63, 32], objectness scores [14, 69, 42, 81], box regression [40, 17], and appearance transfer [57, 34, 52, 5]. This setting is also referred to as domain adaptation, transfer learning, or mixed supervision object detection.

Our goal is to enable streamlined knowledge transfer from a large source dataset to a variety of target domains, so we leverage the objectness scores setting by using the proposals and scores of a pretrained class-agnostic model in WSOD. We primarily compare our method to Zhong *et al*. [81], the state-of-the-art Faster R-CNN approach in this setting. Besides the architecture, another key difference is that [81] allows multiple refinements over the source dataset, while our method only needs one pretraining iteration.

Detection Transformer (DETR). In comparison to CNNs, Transformer-based methods [70] require more data and training time, but excel in the large-data regime and are particularly effective at transfer learning [18]. DETR [6] introduces a Transformer framework for object detection problems. In the DETR pipeline, image features are extracted with a ResNet [28] backbone before being passed into a Transformer encoder. The decoder takes as input a number of learned object queries and attends to the encoder output, from which a feedforward network produces box and class predictions. Unlike Faster R-CNN [55], DETR is trained with a set prediction loss and does not require non-maximum suppression, spatial anchors, or RoI pooling.

The community has made significant progress improving DETR with faster convergence [82, 24, 11] and pretraining tasks [12, 3]. We extend Deformable DETR [82], which uses an efficient attention mechanism for $10\times$ faster training, and DETReg [3], which uses unsupervised pretraining to improve downstream localization performance.

To the best of our knowledge, we propose the first MIL-based DETR variant amenable to WSOD tasks. Notably, Chen *et al*. [7] extend DETR to weak point-based annotations, which is a different setting that does not involve MIL.

3. Weakly Supervised Detection Transformer

We propose the Weakly Supervised Detection Transformer (WS-DETR), which integrates DETR with MIL for scaling WSOD finetuning on novel objects. In contrast to prior work requiring multiple rounds of box refinement [65] or pseudo ground truth mining [81], WS-DETR simplifies the process, requiring only pretraining and a single round of MIL training. Like other Transformer-based methods, WS-DETR is particularly scalable to large pretraining datasets. For WSOD applications, this enables a more accurate understanding of objectness, resulting in higher performance during knowledge transfer to novel objects.

3.1. Class-Agnostic DETR

During pretraining, we utilize a class-agnostic DETR trained on the binary object labels of the source dataset to predict bounding box proposals and objectness confidence scores for use during WSOD finetuning. The class-agnostic DETR model extends Deformable DETR [82], a variant of the vanilla DETR [6] with multi-scale features and improved training efficiency, and predicts a fixed-set size of N object proposals. We use $N = 300$, the default for Deformable DETR. After the Transformer features are computed, a 3-layer network with ReLU returns proposal coordinates $\{\boldsymbol{p}_j\}_{j=1}^N$ and a fully-connected layer returns classification logits $\{s_j\}_{j=1}^N$ interpreted as objectness scores.

During weakly supervised finetuning, the fully-connected layer is dropped in favor of two C-class layers for classification and detection, respectively.

Optionally, we can additionally train a fully supervised class-aware DETR (*i.e.,* using supervised class labels instead of binary object labels) on the source dataset. If so, the class-agnostic model is used for object proposals and scores, while the class-aware model is used for initializing the weakly supervised branch. This strategy begets a performance boost (discussed in Section 4.5) and is convenient as many pretrained models are class-aware to begin with. On the other hand, using a single class-agnostic model as both proposal generator and initialization halves computation.

A key difference between our WS-DETR and previous WSOD models based on Faster R-CNN [55, 81] is that DETR learns end-to-end using a positional embedding – also called object queries – instead of using RoI pooling. Thus, we freeze both the object queries and box prediction

head of the class-agnostic model when finetuning the pretrained DETR. If using a class-aware checkpoint, the object queries are set equal to those of the class-agnostic model; hence, the class-aware Transformer decoder attends to the same locations as its class-agnostic counterpart.

3.2. MIL Classifier

As illustrated in Figure 1, the MIL classifier receives the classification logits $\mathbf{C} \in \mathbb{R}^{N \times C}$ and detection logits $\mathbf{D} \in \mathbb{R}^{N \times C}$ and converts them to an image-level classification prediction. The classification logits are softmaxed over the class dimension (columns), while the detection logits are softmaxed over the detection dimension (rows). Let σ denote the softmax operation for $\boldsymbol{z} \in \mathbb{R}^N$:

$$\sigma(\boldsymbol{z})_i = \frac{\exp z_i}{\sum_{j=1}^N \exp(z_j)}. \tag{1}$$

We define the class-wise and detection-wise softmaxes as $\sigma_{ij}^c(\mathbf{A}) = \sigma((\mathbf{A}^\top)_j)_i$ and $\sigma_{ij}^d(\mathbf{A}) = \sigma(\mathbf{A}_i)_j$ where $\mathbf{A}_i$ is the ith row of $\mathbf{A}$. The softmaxed matrices in the MIL classifier are $\sigma^c(\mathbf{C})$ and $\sigma^d(\mathbf{D})$; they are then multiplied element-wise and summed over the detection dimension to obtain the image-level classification predictions $\{\hat{y}_j\}_{j=1}^C$:

$$\hat{y}_j = \sum_{i=1}^N \sigma_{ij}^c(\mathbf{C})\sigma_{ij}^d(\mathbf{D}). \tag{2}$$

Since the rows of $\sigma^c(\mathbf{C})$ and the columns of $\sigma^d(\mathbf{D})$ are each non-negative and sum to one, we have $\hat{y}_i \in (0, 1)$ for all i. Then, the image-level labels $\{y_j\}_{j=1}^C$ are used to compute the negative log-likelihood loss:

$$\mathcal{L}_{\text{MIL}} = -\frac{1}{C}\sum_{j=1}^C y_j \log(\hat{y}_j) + (1 - y_j)\log(1 - \hat{y}_j). \tag{3}$$

The state-of-the-art method for knowledge transfer from a pretrained class-agnostic model is the objectness regularization technique of [81], which uses the class-agnostic model's objectness scores to guide the training of the detection branch. Let $S(x) = 1/(1 + e^{-x})$ denote the sigmoid operation for $x \in \mathbb{R}$, then

$$\mathcal{L}_{\text{obj}} = \frac{1}{N}\sum_{i=1}^N \big(\max_{1 \le j \le C} S(\mathbf{D}_{ij}) - S(s_i)\big)^2. \tag{4}$$

Hence, the model loss is $\mathcal{L} = \mathcal{L}_{\text{MIL}} + \lambda\mathcal{L}_{\text{obj}}$ for a coefficient λ. During inference, WS-DETR returns box $\boldsymbol{p}_i$ with class prediction and confidence determined by $\text{argmax}_{1 \le j \le C} \sigma_{ij}^c(\mathbf{C})\sigma_{ij}^d(\mathbf{D})$.

3.3. Joint Probability Estimation

We show in Section 4.2 that the objectness regularization technique of [81] is insufficient for general WSOD, as it can suffer from the common MIL weakness of overfitting to distinctive classification features [65]. To rectify this and more effectively utilize the pretrained knowledge from DETR, we propose a formulation for the MIL classifier based on the joint object and class probabilities for each proposal [54]. For a given proposal i, let $c_i = \max_{1 \le j \le C} \sigma_{ij}^c(\mathbf{C})$ and $d_i = \max_{1 \le j \le C} S(\mathbf{D}_{ij})$ be its maximum classification and detection probabilities, respectively.

There are two scenarios which can cause this overfitting problem. First, if c_i is high for a particular class j but $\mathbf{D}_{ij}$ is low, the model may take a penalty in $\mathcal{L}_{\text{obj}}$ to increase $\mathbf{D}_{ij}$ and easily minimize $\mathcal{L}_{\text{MIL}}$ for the image. However, this can be avoided by increasing λ. The more likely explanation is based on a weakness of the objectness regularizer: for a given proposal i, the regularizer only cares about the value of d_i and not whether its position in the row actually lines up with c_i – that is, whether

$$\underset{1 \le j \le C}{\text{argmax}}\, \sigma_{ij}^c(\mathbf{C}) = \underset{1 \le j \le C}{\text{argmax}}\, S(\mathbf{D}_{ij}). \tag{5}$$

If these values are mismatched, this failure case would result in low confidences for every proposal and essentially sort them by c_i, causing overfitting. Indeed we observe that when using our WS-DETR with objectness regularization, the final confidences are typically 0.01 or lower.

We desire the overall probability of these distinctive features to be diminished by a low objectness probability, since the pretrained model should recognize that the feature does not represent an entire object. Thus, we compute the probability $\mathbb{P}[i^{th} \text{ proposal is an object and instance of class } j] = \sigma_{ij}^c(\mathbf{C})S(s_i)$. With this formulation, it is required that a certain proposal should have both a nontrivial classification and objectness probability for it to be included in the final prediction. Using normalized probabilities via softmax [4], we obtain the new image-level classification prediction

$$\hat{y}_j = \sum_{i=1}^N \sigma_{ij}^c(\mathbf{C})\sigma(\boldsymbol{s})_i. \tag{6}$$

Note that our technique is mutually exclusive with objectness regularization. We show in Section 4.2 that our modification successfully prevents overfitting to distinctive classification features. Further, our technique simplifies the network as we are essentially using the objectness confidences from the pretrained DETR in place of a learnable detection branch in the MIL classifier, and it improves convergence by minimizing $\mathcal{L}_{\text{MIL}}$ without $\mathcal{L}_{\text{obj}}$.

3.4. Sparsity in the MIL Classifier

The objectness knowledge present in the pretrained DETR can also be leveraged to reduce noise during multiple instance learning – while there are a fixed number of $N = 300$ proposals, the model typically only detects a few

Table 1: Class-agnostic performance of Faster R-CNN (used by [81]) and DETR methods trained on FSOD-800 and evaluated on each FSOD-200 test split, ignoring classes. We use the codebase of [81], which does not report precision for this task.

Method	mAP	AP50	mAR
Zhong *et al.* [81]	–	–	50.5 ± 2.1
Class-Aware DETR	18.4 ± 1.0	26.9 ± 0.94	62.3 ± 3.0
Class-Agnostic DETR	$\mathbf{30.6 \pm 1.6}$	$\mathbf{43.0 \pm 1.6}$	$\mathbf{65.5 \pm 3.2}$

Table 2: WSOD performance on FSOD-200 splits with FSOD-800 pretraining. Our WS-DETR is initialized with class-agnostic proposal generator and class-aware weights. The supervised DETR is finetuned from the class-aware FSOD-800 checkpoint.

Method	mAP	AP50	mAR
Zhong *et al.* [81]	20.6 ± 0.76	32.7 ± 2.0	34.4 ± 0.43
WS-DETR Base	13.9 ± 1.6	20.0 ± 1.9	60.1 ± 2.4
WS-DETR Sparse	28.5 ± 0.86	$\mathbf{38.5 \pm 0.63}$	$\mathbf{68.0 \pm 4.3}$
WS-DETR Joint	28.6 ± 0.43	37.8 ± 0.87	65.3 ± 1.5
WS-DETR Full	$\mathbf{28.6 \pm 0.25}$	38.2 ± 1.1	67.4 ± 3.9
Supervised DETR	47.7 ± 1.3	64.0 ± 1.0	76.3 ± 1.2

with high objectness scores. To focus more on these confident proposals, we propose utilizing sparsity along the detection dimension of the MIL classifier.

To do so, we apply the sparsemax function [47] instead of the softmax function along the detection dimension in the MIL classifier; this operation zeros out some low-confidence boxes, increasing emphasis on correct classification of likely proposals. Specifically, sparsemax returns the Euclidean projection of a vector $\boldsymbol{z} \in \mathbb{R}^N$ onto the $(N-1)$-dimensional probability simplex $\Delta^{N-1} = \{\boldsymbol{p} \in \mathbb{R}^N : \mathbf{1}^\top \boldsymbol{p} = 1, \boldsymbol{p} \geq \mathbf{0}\}$:

$$\text{sparsemax}(\boldsymbol{z}) = \underset{\boldsymbol{p} \in \Delta^{N-1}}{\arg\min} \|\boldsymbol{p} - \boldsymbol{z}\|_2^2. \tag{7}$$

We then substitute sparsemax$(\mathbf{D}_i)_j$ for $\sigma_{ij}^d(\mathbf{D})$ in Equation 2 and sparsemax$(\boldsymbol{s})_i$ for $\sigma(\boldsymbol{s})_i$ in Equation 6.

While there are many sparsity techniques, we choose sparsemax because of its theoretical justification, its ease-of-use with no hyperparameter tuning, and its successful application in previous MIL architectures [76] (though not in WSOD). The structure of the loss function also makes it particularly well-suited to the MIL problem. In a traditional classification setting, it is possible to get a $\log(0)$ in the loss function because sparsemax can send labels to a probability of exactly zero – Martins and Astudillo [47] define a new sparsemax loss to deal with this issue. But because we multiply sparsemax$(\mathbf{D})$ element-wise with $\sigma^c(\mathbf{C})$ whose entries are > 0, there is some element > 0 in each column of the product. Thus, we still have $\hat{y}_i \in (0, 1)$ for all i, and we can still apply the negative log-likelihood loss.

4. Experiments

4.1. Large-Scale Novel Object Detection

To evaluate the performance of our WS-DETR on highly novel objects, we utilize the the Few-Shot Object Detection (FSOD) dataset [21], designed to test the generalization performance of few-shot learning models on novel objects in a diverse setting. Fan *et al.* [21] constructed the dataset from existing large-scale supervised datasets ILSVRC [58] and Open Images [39], and merged their semantic trees pursuant to Open Images superclasses. The FSOD dataset comprises 1000 classes, of which 800 are reserved for training

and 200 for testing – we call these datasets FSOD-800 and FSOD-200. This train/test split is generated such that the test classes have the largest distance from existing training categories in the semantic tree, enabling a challenging setting for generalization to truly novel objects.

In contrast to few-shot object detection, WSOD requires a target dataset of novel objects for model finetuning. Thus, we utilize FSOD-800 as a source dataset for pretraining, and we create three random train/test splits of FSOD-200 for training and evaluation on novel objects, which will be released for reproducibility. We report the mean and 95% confidence interval of the metrics against each split based on a t-distribution with two degrees of freedom. FSOD-800 has 52,350 images with 147,489 boxes, while the FSOD-200 splits each have 11,322 training images with between 28,008 and 28,399 boxes and 2,830 testing images with between 6,703 and 7,094 boxes. This setting has $4\times$ the source classes and $2\times$ the target classes than the largest datasets used for WSOD previously [58, 69, 81].

While some WSOD methods use the Correct Localization (CorLoc) metric [13] for evaluating localization accuracy, this metric is too lenient as it only requires localizing a single object per image. Hence, we instead use the mean average recall (mAR) at 100 detections per image for comparing class-agnostic proposal quality, though we also report mean average precision (mAP) and AP50 for comparison. In Table 1 we compare the performance of the class-agnostic and class-aware DETR versus the class-agnostic Faster R-CNN of [81] trained on FSOD-800 and evaluated on each FSOD-200 test split. For the class-aware DETR, we ignore the class predictions and evaluate the boxes only. Both DETR variants outperform the Faster R-CNN, and the class-agnostic DETR and class-aware DETRs have similar recall, though the class-agnostic DETR has much better precision. We show in Section 4.5 that this precision improvement translates to superior WSOD performance, justifying the extra pretraining of a class-agnostic model.

For brevity, we introduce short names for each permutation of WS-DETR with our techniques from Sections 3.3 and 3.4. "Base" refers to our model with the objectness reg-

(a) WS-DETR Base   (b) WS-DETR Sparse

(c) WS-DETR Joint   (d) WS-DETR Full

Figure 2: Visualization of how our joint probability technique prevents overfitting to distinctive classification features on the FGVC-Aircraft dataset. Best viewed electronically and zoomed in. Models (c) and (d) both use our technique. The plotted box is the highest confidence detection.

ularization of [81]; "Sparse" refers to ours with sparsity and objectness regularization; "Joint" refers to ours with joint probability estimation only; and "Full" refers to ours with sparsity and joint probability estimation.

We then train our WS-DETR with class-agnostic proposal generator and class-aware weights initialization on each FSOD-200 split. In Table 2, we detail the performance of our model against the state-of-the-art baseline of [81] and a supervised DETR upper bound. The addition of either our joint probability technique or sparsity boosts mAP by nearly 15 points over WS-DETR Base, achieving a new state-of-the-art performance by 8 mAP. This suggests that, for WS-DETR, regularization by itself is insufficient for transferring objectness knowledge learned from the source dataset to the downstream WSOD task. In particular, as we show in the next section, our joint probability technique is critical to prevent overfitting to distinctive classification features. Additionally, while the method of [81] loses 15 mAR during weakly supervised training, our WS-DETR gains 2.5 mAR relative to the class-agnostic pretrained model.

4.2. Joint Probability Prevents Overfitting

The FGVC-Aircraft dataset [45] comprises 10,000 images of 100 types of aircraft whose visual characteristics may differ only slightly between classes. It poses a fairly simple detection problem, as the target objects are large and centered. Additionally, since "airplane" is one of the FSOD-800 source classes, one would expect WSOD models to perform well on this task. We show that our joint probability formulation achieves this outcome, while the objectness regularization technique utilized in previous

Table 3: WSOD performance on the FGVC-Aircraft dataset with FSOD-800 pretraining. The models using our joint probability technique achieve near-supervised performance, while the objectness regularization methods underperform due to overfitting to distinctive classification features.

Method	mAP	AP50	mAR
Zhong *et al.* [81]	14.8	28.7	30.5
WS-DETR Base	5.2	8.5	63.4
WS-DETR Sparse	50.6	57.4	93.2
WS-DETR Joint	77.7	83.6	93.4
WS-DETR Full	**79.1**	**85.0**	**94.2**
Supervised DETR	87.1	88.7	97.9

work [81] limits detection performance. In particular, previous models overfit to distinctive classification features – here, the nose or tail of the aircraft – a weakness observed by [65] and whose remedy has been the subject of several WSOD studies [65, 75, 33]. These solutions typically involve multiple iterative rounds of box refinement via self-training. In contrast, our method leverages the objectness knowledge from the pretrained model to identify the correct proposal without any extra computation.

In Figure 2, we visualize the advantage of our WS-DETR method and how it properly selects the bounding box covering the entire aircraft, while the objectness regularization technique overfits to distinctive features. In Table 3, we display the precision and recall of each model on the FGVC-Aircraft test set and demonstrate that our model achieves near-supervised level performance.

4.3. Finetuning on 2,854 Fine-grained Classes

A practical application for WSOD not captured by current settings is on datasets with many classes which require domain-specific knowledge for labeling. An exemplar of this variety is the iNaturalist 2017 dataset [30], a fine-grained species dataset of 500K boxes and 5,000 classes, 2,854 of which have detection annotations. In fact, Horn *et al.* [30] remark that the bounding box labeling was particularly difficult – since only a domain expert can distinguish between so many different species, they asked labelers to demarcate superclasses only, then backfilled the bounding boxes with the (sometimes incorrect) image-level classification labels. The most recent iNaturalist dataset contains 10,000 species and 2.7 million images, which is nearly impossible to label for detection tasks and represents a highly practical scenario for a weakly supervised approach.

The WS-DETR model has seen different types of plants and animals during pretraining on FSOD-800, but nowhere near the granularity and diversity of iNaturalist with its thousands of leaf classes. This makes iNaturalist an interesting setting for studying knowledge transfer during WSOD training. In Table 4, we detail the performance of our WS-DETR against a state-of-the-art model [81] on the

Table 4: WSOD performance on the iNaturalist 2017 dataset with FSOD-800 pretraining. Our WS-DETR is initialized with class-agnostic proposal generator and class-aware weights. The supervised DETR upper bound is finetuned from the same class-aware FSOD-800 checkpoint. The method of [81] did not converge for the subclasses task.

Method	13 Superclasses			2,854 Subclasses		
	mAP	AP50	mAR	mAP	AP50	mAR
Zhong *et al.* [81]	44.1	76.7	57.1	—	—	—
WS-DETR Base	0.2	0.4	31.9	1.7	3.7	26.3
WS-DETR Sparse	**61.1**	**79.3**	83.3	30.4	38.2	**77.6**
WS-DETR Joint	54.8	70.0	**84.3**	22.1	29.8	75.5
WS-DETR Full	60.7	78.7	83.1	**35.4**	**43.5**	75.5
Supervised DETR	79.2	93.6	88.6	51.5	58.8	85.6

Table 5: WSOD performance on PASCAL VOC 2007 with COCO-60 pretraining. The supervised DETR is finetuned from the same COCO-60 checkpoint. The result of Zhong *et al.* [81] includes pseudo ground truth mining.

Method	mAP	AP50	mAR
WSDDN [4]	—	34.8	—
CASD [33]	—	56.8	—
Zhong *et al.* [81]	—	**59.7**	—
WS-DETR Base	18.2	28.4	58.4
WS-DETR Sparse	24.2	36.5	57.7
WS-DETR Joint	23.4	33.8	58.4
WS-DETR Full	23.6	34.2	57.6
Supervised DETR	55.3	77.3	72.7

13 superclasses and 2,854 subclasses in the dataset. While the method of [81] did not converge on the subclasses, our model achieves 75% of supervised performance.

The iNaturalist experiment reveals several intriguing phenomena. First, our WS-DETR does not converge without our joint probability technique or sparsity, suggesting that our techniques improve training stability as well as performance. Second, the addition of sparsity to our joint probability technique improves results by up to 8.3 mAP, showing its versatility and effectiveness even without a fully-connected detection branch. Third, our WS-DETR outperforms the state-of-the-art by 17 mAP but less than 3 AP50; this indicates that, while Faster R-CNN is able to localize objects at a modest IoU threshold, the improved localization and knowledge transfer in our WS-DETR can have a significant impact on high-precision WSOD performance.

We note that, while WS-DETR Sparse sometimes attains the top performance, we still recommend WS-DETR Full for implementation in practice. WS-DETR Full consistently scores within 0.5 mAP of the best model, and without prior knowledge as to the distinctive features of the target dataset, WS-DETR Sparse may overfit due to its reliance on objectness regularization. Hence, as seen in FGVC-Aircraft and the iNaturalist subclasses, WS-DETR Full greatly outperforms WS-DETR Sparse in fine-grained datasets.

4.4. COCO-60 to VOC Performance

Though the presence of many classes in the source dataset is posited to be essential for effective transfer [69], previous WSOD methods are typically designed for and trained on datasets with few classes and small image sets. One such setting is PASCAL VOC [20] (20 classes), where knowledge transfer methods use COCO-60 [43, 40] (60 classes) for pretraining with no class overlap.

As stated in Section 1, reliance on COCO-60/VOC has drawbacks which limit the usage of previous WSOD models in practice. Yet, for completeness we have tested our method on this common test case (detailed in Table 5). Our best approach is below the leading method [81]. This

is at odds with our performance relative to [81] for diverse datasets with hundreds of novel objects such as FSOD and fine-grained datasets such as iNaturalist and FGVC-Aircraft. We believe this inconsistency illustrates the benefits of our method, which can leverage large-scale pretraining for weakly supervised detection on complex datasets that are common in real-world scenarios. Our results suggest this benefit is due to the scalability of our model and our novel method of finetuning an end-to-end detection model instead of a classification model such as ResNet [28, 81]; however, a limitation of our model – shared with all Transformer methods – is its requirement of a large pretraining dataset for effective downstream performance.

Previous methods work well for the small-scale COCO-60/VOC case but not large and diverse datasets, which we believe are more common in real-world applications. Together with our scaling study in Section 4.5, this shows that it is time for WSOD research to move beyond what appears to be an over-optimization to COCO-60/VOC, which is not a useful analogue for real-world datasets, and address the complex datasets that we study in our work.

4.5. Ablation Study

We perform a scaling study on FSOD-800 pretraining with WS-DETR Full and find that class quantity contributes more to downstream WSOD performance than image quantity (see Figure 3). Group 1, the solid lines in the figure, are a random split of FSOD-800 with all classes represented. Group 2, the dashed lines in the figure, have the same number of images as Group 1 but with that same proportion of classes. This experimental setup isolates the effect of increased pretraining classes with the same number of total images. We take 3 random splits of FSOD-800 at each percentage level for each group and finetune on the 3 splits of FSOD-200. We report the mean and 95% CI with respect to a t-distribution with 8 *dof*. This is the first rigorous testing and proof of the hypothesis of Uijlings *et al.* [69] that class quantity is more important than image quantity for WSOD
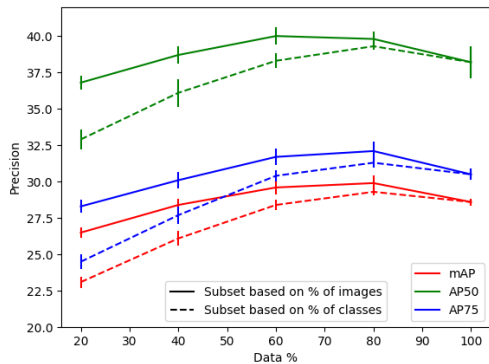
Figure 3: Scaling study of FSOD-800 pretrained WS-DETR Full with FSOD-200 WSOD. We test pretraining with a percentage of images vs. a percentage of classes, then perform WSOD training and evaluate on our held-out test set. This shows pretraining class quantity contributes more to performance than image quantity.
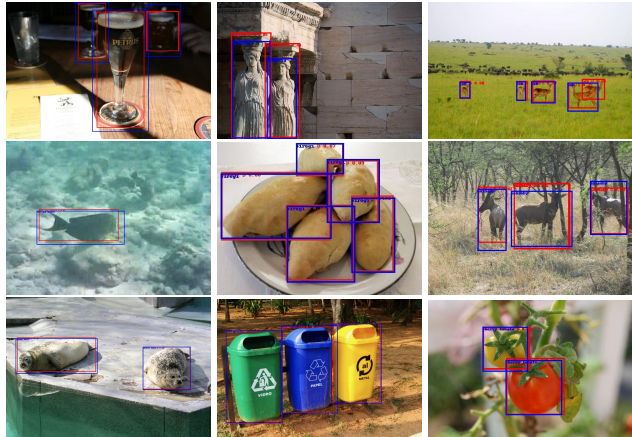

Figure 4: Sample of results of WS-DETR Full on the FSOD-200 held-out test set. Best viewed electronically and zoomed in. Blue represents ground-truth and red represents our WS-DETR prediction. The model has not seen bounding box labels for classes in the test set.

Table 6: WSOD performance on each FSOD-200 split with FSOD-800 pretraining. We utilize our joint probability technique and no sparsity. The class-aware DETR is pretrained on FSOD-800 with class labels, while the class-agnostic DETR is pretrained with only binary object labels.

Proposal Generator	Weights Init.	mAP	AP50
Aware	Agnostic	18.0 ± 1.1	24.3 ± 1.4
Aware	Aware	22.1 ± 1.7	29.7 ± 2.3
Agnostic	Agnostic	27.0 ± 1.0	35.8 ± 1.7
Agnostic	Aware	$\mathbf{28.6 \pm 0.43}$	$\mathbf{37.8 \pm 0.87}$

pretraining, and it justifies our usage of FSOD-800 in place of a larger dataset with less classes such as COCO [43]. The lowest proportion of classes we test (160 classes) is still nearly $3\times$ that of COCO-60 [43, 40]; the performance gap at this level suggests that standard datasets used for WSOD pretraining are an order of magnitude too small.

We note that Figure 3 peaks around 80% of data; we believe this is due to the random dropout of irrelevant classes from the pretraining dataset. A future direction is to quantify the impact of class diversity on WSOD performance; indeed, we and Uijlings *et al*. [69] use quantity as a proxy for diversity, which has been shown in theoretical works [19] to be necessary for effective finetuning on novel classes.

In our above experiments, we used a class-agnostic pretrained DETR as the proposal generator and a class-aware pretrained DETR as the weights initialization. This requires pretraining two separate DETR models, which can be computationally intensive for large source datasets. In these cases, we can halve computation by using the class-agnostic model as the weights initialization. In Table 6, we detail

the performance of different proposal generator and weights initialization strategies on each FSOD-200 split. The WS-DETR trained with class-agnostic proposal generator and weights initialization only loses 1.6 mAP and 2.0 AP50 compared to the best model; this suggests that while the feature space learned during the class-aware pretraining is a useful initialization, the class-agnostic model can learn most necessary features during WSOD finetuning.

5. Conclusion

We propose the Weakly Supervised Detection Transformer (WS-DETR), which integrates DETR with an MIL architecture for scalable WSOD finetuning on novel objects. Our hybrid model leverages the strengths of both two-stage detectors and the end-to-end DETR framework. In comparison to existing WSOD approaches, which only operate at small data scales and require multiple rounds of training and refinement, our WS-DETR utilizes a single pretrained model for knowledge transfer to WSOD finetuning in a variety of practical domains. We achieve state-of-the-art performance in novel and fine-grained settings, and our scaling study reveals that class quantity is more important than image quantity for WSOD pretraining.

Potential negative social impact. Object detection models have malicious potential for surveillance. Ours could have unintended negative impact by lessening the labeling needed to detect fine-grained categories of people; we explicitly discourage these applications. Additionally, the environmental cost of pretraining Transformers on massive datasets is significant, so we will release our checkpoints for others to utilize with minimal added emissions.

References

[1] Baptiste Angles, Yuhe Jin, Simon Kornblith, Andrea Tagliasacchi, and Kwang Moo Yi. MIST: Multiple instance spatial transformer. In *34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. DETReg: Unsupervised pretraining with region priors for object detection. In *35th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Tianyue Cao, Lianyu Du, Xiaoyun Zhang, Siheng Chen, Ya Zhang, and Yan-Feng Wang. CaT: Weakly supervised object detection with category transfer. In *18th International Conference on Computer Vision (ICCV)*, 2021.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *16th European Conference on Computer Vision (ECCV)*, 2020.

[7] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *7th International Conference on Learning Representations (ICLR)*, 2019.

[9] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian sheng Hua. Slv: Spatial likelihood voting for weakly supervised object detection. In *33rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic DETR: End-to-end object detection with dynamic attention. In *18th International Conference on Computer Vision (ICCV)*, 2021.

[12] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: Unsupervised pre-training for object detection with transformers. In *34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[13] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *11th European Conference on Computer Vision (ECCV)*, 2010.

[14] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision (IJCV)*, 100(1):275–293, 2012.

[15] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

[17] Bowen Dong, Zitong Huang, Yuelin Guo, Qilong Wang, Zhenxing Niu, and Wangmeng Zuo. Boosting weakly supervised object detection via learning bounding box adjusters. In *18th International Conference on Computer Vision (ICCV)*, 2021.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations (ICLR)*, 2021.

[19] Simon S. Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *9th International Conference on Learning Representations (ICLR)*, 2021.

[20] Mark Everingham, Luc Van Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[21] Qi Fan, Wei Zhuo, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *33rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Carolina Galleguillos, Boris Babenko, Andrew Rabinovich, and Serge Belongie. Weakly supervised object localization with stable segmentations. In *10th European Conference on Computer Vision (ECCV)*, 2008.

[23] Mingfei Gao, Ang Li, Ruichi Yu, Vlad I. Morariu, and Larry S. Davis. C-WSL: Count-guided weakly supervised localization. In *15th European Conference on Computer Vision (ECCV)*, 2018.

[24] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of DETR with spatially modulated co-attention. In *18th International Conference on Computer Vision (ICCV)*, 2021.

[25] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. TS-CAM: Token semantic coupled attention map for weakly supervised object localization. In *18th International Conference on Computer Vision (ICCV)*, 2021.

[26] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-MIDN: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *17th International Conference on Computer Vision (ICCV)*, 2019.

[27] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. LSDA: Large scale detection through adaptation. In *17th Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[30] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The iNaturalist challenge 2017 dataset. In *31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(4):814 – 830, 2015.

[32] Luwei Hou, Yu Zhang, Kui Fu, and Jia Li. Informative and consistent correspondence mining for cross-domain weakly supervised object detection. In *34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[33] Zeyi Huang, Yang Zou, Vijayakumar Bhagavatula, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. In *34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[34] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[36] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *14th European Conference on Computer Vision (ECCV)*, 2016.

[37] Satoshi Kosugi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object-aware instance labeling for weakly supervised object detection. In *17th International Conference on Computer Vision (ICCV)*, 2019.

[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[39] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128(1):1956–1981, 2020.

[40] Seungkwan Lee, Suha Kwak, , and Minsu Cho. Universal bounding box regression and its applications. In *14th Asian Conference on Computer Vision (ACCV)*, 2018.

[41] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *17th International Conference on Computer Vision (ICCV)*, 2019.

[42] Yan Li, Junge Zhang, Kaiqi Huang, and Jianguo Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(3):639 – 653, 2019.

[43] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *13th European Conference on Computer Vision (ECCV)*, 2014.

[44] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *14th European Conference on Computer Vision (ECCV)*, 2016.

[45] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[46] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *10th Conference on Neural Information Processing Systems (NeurIPS)*, 1998.

[47] André F. T. Martins and Ramón F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *33rd International Conference on Machine Learning (ICML)*, 2016.

[48] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *18th International Conference on Computer Vision (ICCV)*, 2021.

[49] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. Unveiling the potential of structure preserving for weakly supervised object localization. In *34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[50] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *16th International Conference on Computer Vision (ICCV)*, 2017.

[51] Yongri Piao, Jian Wang, Miao Zhang, and Huchuan Lu. MFNet: Multi-filter directive network for weakly supervised salient object detection. In *18th International Conference on Computer Vision (ICCV)*, 2021.

[52] Amir Rahimi, Amirreza Shaban, Thalaiyasingam Ajanthan, Richard Hartley, and Byron Boots. Pairwise similarity knowledge transfer for weakly supervised object localization. In *16th European Conference on Computer Vision (ECCV)*, 2020.

[53] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *18th Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

[56] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang andMing Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *33rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[57] Mrigank Rochan and Yang Wang. Weakly supervised localization of novel objects using appearance transfer. In *28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[59] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Jing Li, Xiangyu Zhang, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *17th International Conference on Computer Vision (ICCV)*, 2019.

[60] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Yongjian Wu, and Feiyue Huang. UWSOD: Toward fully-supervised-level capacity weakly supervised object detection. In *34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[61] Yunhang Shen, Rongrong Ji, Yan Wang, Zhiwei Chen, Feng Zheng, Feiyue Huang, and Yunsheng Wu. Enabling deep residual networks for weakly supervised object detection. In *16th European Conference on Computer Vision (ECCV)*, 2020.

[62] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[63] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *16th International Conference on Computer Vision (ICCV)*, 2017.

[64] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. PCL: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(1):176–191, 2018.

[65] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[66] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[67] Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Visual and semantic knowledge transfer for large scale semi-supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12):3045 – 3058, 2018.

[68] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(1):154–171, 2013.

[69] Jasper R. R. Uijlings, S. Popov, and V. Ferrari. Revisiting knowledge transfer for training object class detectors. In *31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[71] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[72] Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S. Kevin Zhou, and Shuguang Cui. Shallow feature matters for weakly supervised object localization. In *34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[73] Jinheng Xie, Cheng Luo, Xiangping Zhu, Ziqi Jin, Weizeng Lu, and Linlin Shen. Online refinement of low-level feature based activation map for weakly supervised object localization. In *18th International Conference on Computer Vision (ICCV)*, 2021.

[74] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *17th International Conference on Computer Vision (ICCV)*, 2019.

[75] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. WSOD^2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *17th International Conference on Computer Vision (ICCV)*, 2019.

[76] Lijun Zhang, Srinath Nizampatnam, Ahana Gangopadhyay, and Marcos V. Conde. Multi-attention networks for temporal localization of video-level labels. In *3rd Workshop on YouTube-8M Large Scale Video Understanding, 17th International Conference on Computer Vision (ICCV)*, 2019.

[77] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[78] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *15th European Conference on Computer Vision (ECCV)*, 2018.

[79] Xiaopeng Zhang, Yang Yang, and Jiashi Feng. Learning to localize objects with noisy labeled instances. In *33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[80] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[81] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *16th European Conference on Computer Vision (ECCV)*, 2020.

[82] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations (ICLR)*, 2021.

[83] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *13th European Conference on Computer Vision (ECCV)*, 2014.