Instance-Dependent Noise Refinement in Segment Anything Model for Weakly Supervised Object Detection

Fariborz Taherkhani[1] and Ehsan Kazemi[2]

Carnegie Mellon University[1], University of California-Davis[2]
fariborztaherkhani@gmail.com[1]

Abstract. We propose a new framework for Weakly Supervised Object Detection (WSOD), a domain that traditionally relies on image-level labels. In addressing the inherent limitations of current WSOD methods, particularly their reliance on image-level annotations that result in inaccurate bounding box selections, we develop a framework that iteratively utilizes weak supervision and refines it to progressively enhance the supervision of the object detector throughout the training process. Specifically, we employ the Segment Anything Model (SAM) to generate initial pseudo-labels bounding boxes from the point prompts generated by Class Activation Mapping (CAM). Our approach tackles the challenge of label noise, where pseudo-labels bounding boxes might only capture parts of objects. We enhance our ability to distinguish between complete and partial detected objects by leveraging an instance-dependent, particularly part-based noise correction model. Our method is inspired by learning methods focusing on part-based representations for object detection and recognition, as well as from human perception, which typically simplifies complex visual information into simpler, constituent parts. Our experiments, conducted in various settings beyond WSOD, including Semi-Supervised Object Detection (SSOD) and Weakly Supervised Instance Segmentation (WSIS), validate the efficacy of our approach.

Keywords: Weakly supervised object detection · Noisy bounding boxes

1 Introduction

Supervised neural networks excel in object detection, a crucial computer vision task, thanks to well-annotated datasets with comprehensive bounding boxes and segmentation details [45]. Nonetheless, when juxtaposed with image classification, the process of annotating objects for detection is notably more resource-intensive and time-consuming [71]. We turn our attention to the domain of WSOD, a technique that endeavors to educate object detectors exclusively utilizing image-level category labels. Previous WSOD models [10,86,54,77] have frequently relied on generating object proposals using a heuristic approach with low precision but high recall [89,108], and subsequently applying multiple instance learning [5,97,96] to recover proposals with a high likelihood. Thus, given

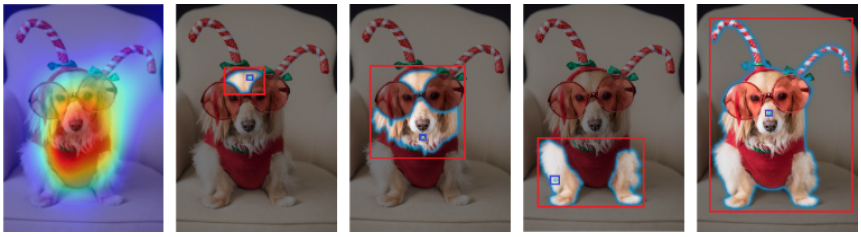


Fig. 1: Pseudo-label bounding boxes from SAM using different point prompts.

the primary difficulties encountered in WSOD stem from precision in locating objects, a strong model such as Segment Anything Model (SAM) [44] which is able to segment or localize any object in an image with high precision, allows the WSOD problem to be recast as an open-set classification problem in a weakly supervised fashion [55,78].

In a recent development, the introduction of SAM and its applications in segmentation [15,34,58,18,65] and detection tasks [102,84] marks a pioneering stride in the realm of image analysis. SAM provides a flexible and dynamic basis for accurately segmenting objects and regions in images. It is uniquely versatile, responding to a range of input types such as points, boxes, masks, or even textual cues. Inspired by SAM model, we revisit WSOD based on pseudo-labeling approaches [50,20,22], wherein the model initially makes predictions on the training data using only image-level labels and essentially guesses the locations of objects. These predictions are not as accurate as true annotations but provide a starting point for training. Our WSOS model begins with a pre-trained classifier using CAM [104] to create attention maps, which are processed to form point prompts for the SAM model. SAM utilizes the provided prompts to create masks and, consequently, bounding boxes. These bounding boxes serve as a form of weak supervision signal, enhancing the performance of our object detection model in a weakly supervised learning manner.

The integration of CAM and SAM in our framework is driven by a goal. CAM highlights the most distinctive parts of an object in images, but it may miss the full object boundaries. SAM, which requires minimal supervision, can use CAM's cues to segment the entire object effectively. However, the pseudo-labels generated by SAM sometimes result in bounding boxes that imperfectly capture only portions of the object, rather than encompassing it entirely. This partial capture introduces noise into the pseudo-labels bounding boxes. Moreover, in some cases, the CAM output is also noisy and extends beyond the object's boundary, leading to SAM inadvertently capturing background portions as well. These issues persist even when sampling multiple point prompts from CAM. As an example, as shown in Fig. 1, SAM partially captures the object based on the point prompts from CAM, with red bounding boxes indicating these areas, while the small blue boxes show the locations of the point prompts from CAM.

To address the challenge of training a network with noisy labels, such as training an object detector with noisy bounding boxes, there are two main categories:

model-free and model-based algorithms [4]. Model-free methods use heuristics, like selecting examples with minimal losses, to reduce label noise effects without directly modeling the noise [99,33]. Although these methods can be effective in practice, they lack guaranteed reliability because they do not explicitly model label noise. In contrast, model-based algorithms are designed to both model and learn from label noise [8,92,72]. These algorithms usually employ a transition matrix $T(x)$ that represents the label noise generation process depending on the instances, where $T_{ij}(x)$ is element of matrix $T(x)$, denoting the probability of an observed noisy label j given the actual label i and the instance x. These methods enable the learning of an optimal classifier from noisy data by using the noisy class posterior and the transition matrix to infer the clean class posterior [8,100]. However, approximating the transition matrix is an ill-posed problem [19]. To simplify and make this process more tractable, researchers often adopt practical assumptions about this matrix, tailoring them to the tasks at hand and the specific characteristics of the noise. These assumptions might include the symmetry of the matrix [67], upper-bounded noise rates for instances [19], or the dependency of an instance's noise solely on the instance's parts [92]. In some cases, it is assumed that the noise is even instance-independent, which is referred to as class-conditional label noise modeling [69,72].

In this work, we proceed with a practical assumption that the noisiness of a bounding box is directly influenced by the elements it contains. Consequently, we approximate instance-dependent label noise by utilizing part-dependent label noise to assess the noise levels of the bounding box instances. Specifically, we approximate the transition matrix for each bounding box by combining the transition matrices of its constituent parts. This method addresses challenges encountered with bounding boxes generated by SAM, which may include irrelevant parts or components (e.g., background rather than the actual object) or omit crucial parts of the object (e.g., capturing only part of it). These factors contribute to the noisiness of the bounding boxes. Our approach is supported not only by various computational theories and learning methods emphasizing the importance of part-based representations in object detection [2,1,28] and recognition [36,9] but also by human perceptual which typically involves breaking down complex visuals information into simpler constituent parts [9,64,90].

2 Related Work

2.1 Weakly Supervised Object Detection (WSOD)

Multiple instance learning-based methods [25] in WSOD treat each image as a bag of instances (e.g., regions) and learns to classify these bags. WSDDN [101] uses a two-stream network for classification and localization, but tends to focus on distinctive object parts. WSOD2 [101] improves accuracy by merging adaptive training for object detection. OICR [86] improves WSDDN by solving the discriminative region issue through a three-step refinement of instance classifiers. PCL [85] enhances OICR by adding a robust proposal generation

component using proposal clustering. SDCN [53] combines segmentation and detection, featuring branches for bounding box detection and segmentation masks. ICMWSD [74] improves upon SDCN by focusing on contextual details rather than just an object's distinctive parts. Additionally, [24,29] use segmentation maps to generate instance proposals with contextual information.

Moreover, **CAM-based methods** similar to WSDDN, often overemphasizes an object's distinctive parts, leading to a discriminative region issue. For example, WCCN [24] tackles the discriminative region issue in object detection with a three-stage cascaded network. CASD [38] further improves the accuracy by incorporating comprehensive attention and self-distillation techniques.

2.2 Weakly Supervised Instance Segmentation (WSIS)

WSIS methods are divided in two main categories: the first category uses instance-level bounding-box annotations to guide segmentation models, using techniques such as box-driven segmentation [20,41], multiple instance learning [37], iterative refinement [52], joint probabilistic objectives [6], and object category density maps [21]. The second category tackle WSIS with image-level labels, exploring class response maps [107,106], localization cue propagation [63,46,3], sequential detection-segmentation integration [79]. Recent advancements in WSIS are mostly moving towards a blend of top-down detection and bottom-up segmentation, simplifying training and improving the performance [80].

2.3 Semi-Supervised Object Detection (SSOD)

SSOD involves training networks using a combination of labeled and unlabeled data [13,39,62,87]. In SSOD, [95,105,93] introduce a consistency-based approach, which ensures that predictions for an input image and its flipped version are consistent. Method [82] suggests employing weak data augmentation for model training and strong data augmentation for generating pseudo-labels. Other approaches such as [61] which presents an "Unbiased Teacher" method to tackle issues related to pseudo-labeling bias, and [93] which proposes a soft teacher mechanism and a box jittering to enhance the detection performance.

2.4 Learning with Noisy Labels

Training CNN with noisy labels methods is a dynamic research area, primarily focused on classification tasks. Techniques developed to manage noisy labels include sample selection [33,40], label correction [66,83], and the use of robust loss functions [31,103]. Recently, this research has extended into the field of object detection, bringing several significant contributions [57,39]. Method [11] explored the effects of different label noise types on object detection, proposing a unique co-teaching method for each object to lessen the noisy label impact. Method [51] crafted a framework that cycles between rectifying label noise and refining the model, specifically focusing on noisy category labels and imprecise bounding boxes. Additionally, method [94] proposed a meta-learning approach to deal with noisy labels, using a limited set of clean data samples.

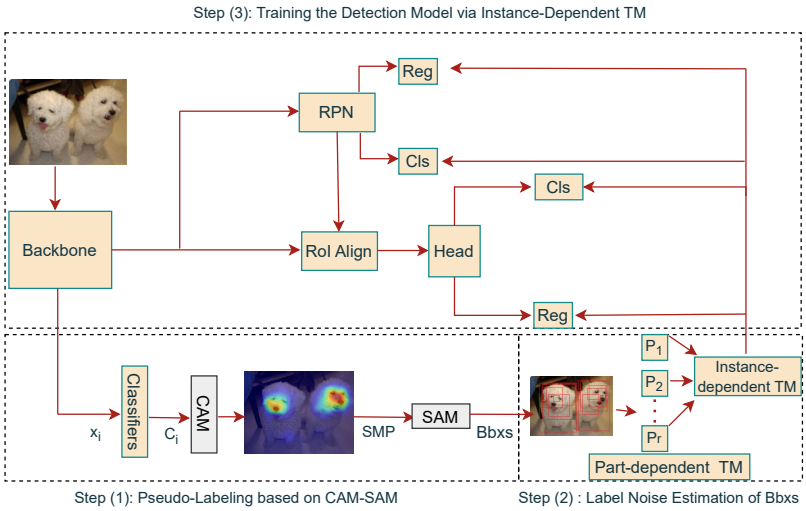


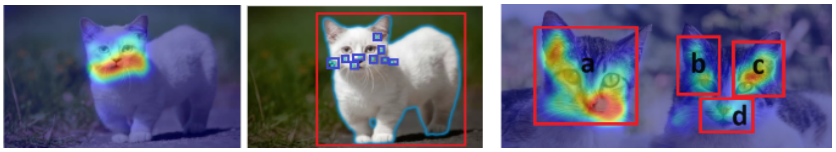
Fig. 2: Instance-Dependent Noise Refinement for Segment Anything Model.

3 Methodology

In this section, we explain our WSOD framework that iteratively utilizes weak supervision and refines it to progressively enhance the supervision of the object detector throughout the training process. The outline of our WSOD model is depicted in Fig. 2. Here, we use a Faster R-CNN-like architecture for detection.

Our WSOD model is a three-step method. Initially, we fine-tune a pre-trained classifier as our backbone with training images and their associated class labels to generate CAMs. These CAMs aid in generating point prompts and facilitate sampling by providing rough estimations of regions likely to indicate object presence. This process involves a post-processing routine where sampling regions are delineated through thresholding on CAMs, allowing for the uniform sampling of point prompts within each identified region. Subsequently, these point prompts individually serve as inputs to the SAM model, which then produces corresponding masks and bounding boxes (Step (1) in Fig. 2 indicates this process).

Given the inherent noisiness in these pseudo-labeled bounding boxes, our method includes a second step to estimate and mitigate this uncertainty. We model the instance-dependent transition matrix for each bounding box based on the transition matrices of its constituent parts, positing that the noise level of an entire instance is directly influenced by the noise characteristics of its individual components (Step (2) in Fig. 2). Consequently, this instance-dependent transition matrix is utilized as a supervisory signal in the third step, guiding the training of the detection model's parameters (Step (3) in Fig. 2). In the subsequent sections, we will delve into the details of each step.



(a) (b)

Fig. 3: Cases where duplicate pseudo-labeled bounding boxes are generated.

3.1 Pseudo-Labeling based on CAM-SAM

Here, we provide a detailed explanation of the first step. In real-world WSOD scenario, we encounter images containing single or multiple objects of various or same classes. This introduces a multi-label classification challenge when training a classifier on such images. While CAM can be adapted for multi-label scenarios to generate points prompts for SAM, the accuracy and quality of CAM may deteriorate, potentially impacting overall performance [43]. To address this, we approach the multi-label problem as a set of binary classification tasks. We employ a pre-trained classifier and fine-tune only the last two fully connected layers for each label. This process ensures CAM produces more precise object regions, a crucial factor for overall performance.

After training a set of binary classifiers, we generate point prompts from distinct activation maps for each class. We apply a 0.5 threshold to CAM values for each class, filtering out irrelevant areas. We then adopt a sampling approach, evenly shifting by 15 pixels both vertically and horizontally from the pixel locations within the segmented areas after thresholding. Due to the noise in CAM outputs post-thresholding, which results in multiple segmented areas appearing on the same object (see patches b, c, d in Fig. 3b, all patches are located on the same object), and the SAM behavior that may produce identical outputs for different point prompts on the same object (See Fig. 3a, where the small blue boxes indicate the locations of the point prompts), we may encounter duplicate bounding boxes in this process, which we then remove. The sequence of the CAM-SAM pseudo-labeling process is depicted in Step (1) of Fig. 2.

3.2 Instance-Based Transition Matrix for Pseudo-Labeled Boxes

Here, we provide a detailed explanation of the second step. Motivated by various learning models that underscore the significance of part-based representations in object detection [2,1,28] and recognition [36,9] , as well as human perceptual processes, which often involve breaking down complex visual information into simpler constituent parts [9,64,90], we aim to approximate instance-dependent label noise of the pseudo-labeled bounding boxes obtained from the first step by leveraging part-dependent label noise of the their constituent parts. Here, we make a practical assumption concerning instance-dependent label noise, positing that an instance's noise is related to its constituent elements. In our case, the

noise usually includes scenarios where certain parts of the object might be missing because the SAM is unable to encompass the entire object. Moreover, noise may stem from the outputs of the CAM or from the SAM itself, particularly when it inadvertently includes portions of the background or extends beyond the object's boundaries. Following the part-based representation approaches [47,2,1,28] where instances (i.e., objects in the bounding boxes) can be approximately reconstructed by a combination of parts, we use the practical and intuitive assumption as presented in [92] for estimating the instance-dependent transition matrices from the combination of part-dependent transition matrices of the object's constituent parts, positing that the parameters for combining these matrices are identical to those employed in the reconstruction of an instance. The logic behind this assumption is that the learned parts carry semantic meaning [47], and their roles in recognizing the instance should mirror their importance in its interpretation and annotation [2,1,28].

Part-based Representation Learning: Non-negative Matrix Factorization (NMF) is a key role for part-based learning approaches, breaking down a positive data matrix into two non-negative matrices [47]. NMF focuses on additive combinations which its adaptability is demonstrated through several variants: Convex-NMF [26] emphasizes data-derived basis vectors; ONMF [98] adds orthogonality for enhanced performance; Semi-NMF [88] accepts data and vectors with both positive and negative values; LCNMF [59] improves robustness by expanding the basis vectors' geometric span; and Truncated CauchyNMF [32] robustly manages outliers by trimming large errors. In this work, we adapt a NMF-based model to learn the parts of the objects.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be the bounding boxes generated in the first step, which may contain noise. Here, y_i is the label that may be either 0 or 1, indicating whether x_i is a noisy bounding box. Meanwhile, $x_i \in \mathbb{R}^d$ represents the normalized representation of an image patch obtained from the SAM using the backbone. The parts-based representations learning can be formulated as:

$$\min_{W_d \in \mathbb{R}^{d\times r}, E(x_i)\in\mathbb{R}_+^r,\, i=1,\ldots,n} \sum_{i=1}^{n} ||x_i - W_d E(x_i)||_2^2, \tag{1}$$

where $||.||_2$ is a Frobenius norm and W_d represents the matrix whose columns correspond to parts of the instances, with each column representing a different part. The term $E(x_i)$ refers to the coefficients that are used to reassemble the instance x_i from these r parts.

In contrast to existing methods [2,1,28,92] for addressing the optimization problem in Eq. 1, our approach harnesses deep learning's power. Unlike method [92], which uses a linear optimization equation and relies on deep representations from a neural network but introduces an additional constraint and consolidates training data into a single matrix—raising scalability concerns for extensive, real-world datasets—we adopt an Auto-Encoder (AE) for optimizing Eq. 1. Moreover, we incorporate non-negativity constraints to preserve data integrity, a critical factor for part-based representation learning. Our approach also introduces

sparsity in the encoder to concentrate on essential components within bounding boxes. These constraints and the sparse representation in our approach which are not present in[92] enhance the ability to accurately represent objects by excluding irrelevant parts [7]. In Eq. 1, $E(x_i)$ represents the AE's encoder output, and W_d denotes the decoder's parameters, with Eq. 1 serving as the AE's reconstruction loss. Notably, in techniques focusing on part-based representations, both input data and decomposed matrices (i.e., $E(X_i)$ and W_d) are typically non-negative. To maintain this non-negativity and enhance our model's part-based representation, we employ ReLU activations in intermediate layers and impose non-negativity constraints on AE's parameters, aligning with practices in Non-negativity Constrained AE (NCA) [49,36,7]. Furthermore, we introduce sparsity in the encoder to ensure that object instances within bounding boxes are reconstructed using minimal yet significant components, effectively omitting irrelevant parts for object representation. Sparsity enforcement is achieved through several methods [35,73], with a common technique involving the restriction of hidden units' activity via KL divergence function [48,68]. Based on the criteria, the objective function for training the AE parameters is:

$$J(W) = J_{\mathrm{AE}}(W) + \alpha J_{KL}(p\|\hat{p}) + \beta f(W), \tag{2}$$

where $W = \{W_e, W_d\}$ are the AE parameters and $J_{\mathrm{AE}}(W)$ is the reconstruction loss defined in Eq. 1, the $J_{KL}(.)$ is the KL term for sparsity constraint and $f(.)$ is a quadratic function for enforcing the non-negativity constrained in AE [70,36,7] where $f(w_i) = w_i^2$ if $w_i < 0$ and $f(w_i) = 0$ if $w_i >= 0$. Here, α and β are the hyperparameters that balance the loss terms. The parameter p, which controls the level of sparsity, is set to a very small number close to 0, while $\hat{p}$ denotes the average activation of a hidden unit over a batch. Let $h_{ij}(x^{(k)})$ be the activation output of the hidden unit in the j th unit of the i th layer and $x^{(k)}$ is the input to it; then $\hat{p}$ is calculated as $\hat{p} = \frac{1}{m}\sum_{k=1}^{m} h_{ij}(x^{(k)})$ using m samples.

Instance-dependent Transition Matrices Estimation: To estimate the transition matrices specific to each instance, it's necessary to determine the transition matrices for each part and the parameters for their combination. Yet, this task faces a challenge because directly decomposing the instance dependent transition matrix into part-dependent matrices and combination parameters is a problem without a clear solution [19]. Following [92], we tackle this problem by presuming that the parameters used to rebuild the instance dependent transition matrix are the same as those used for an instance's reconstruction. This approach is based on the understanding that the parts we learn carry semantic significance [47], and thus, their role in recognizing the instance is assumed to be comparable to their importance in its interpretation and labeling [1,9]. Let $T(x_i)$ be instance transition matrix for sample x_i, then it can be approximated as follows:

$$T(x_i) \approx \sum_{k=1}^{r} E_k(x_i) P_k, \tag{3}$$

where P_k is the k-th part-dependent transition matrix, and $E(x_i)$ is the reconstruction coefficients as defined in Eq. 1. Note that while P_k is unknown, it can be derived from those representations for which we have a high degree of confidence in their transition matrix. These particular samples or representations are identified as anchor points in the literature [60]. Specifically, these representations are considered clean and are presumed to be free of noise. For instance, x_i is an anchor point if $\Pr(y = 1|x_i) = 1$, where y is the label, with $y = 1$ indicating a clean representation and $y = 0$ indicating a noisy one. Consequently, for the anchor sample x_i, the probability $\Pr(y = k|x = x_i)$ is given by $\sum_{j=1}^{c=2} \Pr(y = k|y = j; x = x_i)\Pr(y = j|x = x_i) = T_{1k}(x_i)$. This is justified by the law of total probability and the fact that $\Pr(y = 0|x = x_i) = 0$ and $\Pr(y = 1|x = x_i) = 1$, which means $T_{11}(x) = 1$ and $T_{ij} = 0$ if $(i \neq 1$ and $j \neq 1)$.

We employ an attention mechanism to articulate the representations of anchor points. Specifically, for a given feature vector x_i and its corresponding CAM with u separate parts for the k-th class of objects $(c_{ik}^{(m)})$, we define the element-wise product $x_i \odot \sum_{m=1}^{u} c_{ik}^{(m)}$ as the representation of the anchor point. This method is adopted to accentuate the elements of the representation that are fundamentally indicative of the object class targeted by the CAM.

Using anchor points, we aim to predict the part-dependent transition matrices by aligning them with the corresponding element of the reconstructed instance-dependent transition matrices, that is, $T(x_i) = \sum_{j=1}^{r} E_j(x_i)P_j$. Assume $(x_1^{(k)}, \ldots, x_n^{(k)})$ represents n anchor points for k-th class of c classes of objects. We determine the part-dependent transition matrices for each class of objects by minimizing the reconstruction error. Then, we put forward the following optimization problem to obtain the part-dependent transition matrices:

$$\min_{P_{11},\ldots,P_{cr}} \sum_{i=1}^{c}\sum_{k=1}^{n} \left\| T\left(x_k^{(i)} \odot \sum_{m=1}^{u} c_{ik}^{(m)}\right) - \sum_{j=1}^{r} E_j\left(x_k^{(i)} \odot \sum_{m=1}^{u} c_{ik}^{(m)}\right) P_{ij} \right\|_2^2. \tag{4}$$

3.3 Training

Here, we provide a explanation of the third step. Our approach is straightforward in training step. We begin by optimizing Eq. 2 to determine the parameters of the AE, which subsequently allows us to identify the coefficients $E(x_i)$ as outlined in Eq. 1. Following this, we leverage the anchor representations to learn the instance-dependent transition matrices. Specifically, this process involves with the minimization of Eq. 4, through which we acquire the part-dependent transition matrices. Once these matrices are obtained, we proceed to calculate the instance-dependent transition matrix for each individual instance, as specified by Eq. 3. The final step in our training step involves employing the instance-dependent transition matrices to train the detection model. Specifically, we choose clean bounding boxes by selecting those with the highest $T[1, 1]$ value compared to other elements in the instance-dependent confusion matrix, indicating a high probability of being non-noisy. This process ensures that the model is trained on less noisy pseudo-label bounding boxes.

4 Experiments

4.1 Datasets and Evaluation Metrics

We test our model on three benchmarks: PASCAL VOC 2007, VOC 2012 [27], and MS-COCO [56]. Both PASCAL VOC 2007 and VOC 2012 include 20 object classes plus a background class. Specifically, the VOC 2007 dataset has 9,962 images, split into 2,501 for training, 2,510 for validation, and 4,951 for testing. The VOC 2012 dataset contains 22,531 images, with 5,717 for training, 5,823 for validation, and 10,991 for testing. Our experiments adhere to the typical WSOD setup [10,14,16,23], where training is done on the combined train and validation sets, and testing is done on the test set. Regarding MS-COCO, the dataset used for training includes 82,783 images, and a separate set of 40,000 images is used for testing. During the training phase, only labels at the image level are used. For our evaluation criteria, a predicted bounding box is deemed accurate if its Intersection over Union (IoU) with the actual box is above 0.5, unless specified otherwise. We report the mean Average Precision (mAP) at an IoU threshold of 0.5 for PASCAL VOCs. For MS-COCO, we present mAP at 0.5 and also mAP averaged over IoU thresholds ranging from 0.5 to 0.95 in 0.05 increments.

4.2 Implementation Details

The WSOD backbone used in our experiments is ResNet50, pre-trained on ImageNet. The batch size is set to be 4. We set the maximum iteration numbers to 100k, 1680k and 220k for VOC 2007, VOC 2012, and MS-COCO, respectively. The entire WSOD network undergoes optimization through SGD with a momentum of 0.9, an initial learning rate of 0.001, and a weight decay of 0.0005. Learning rate decay occurs at the 50k th, 80k th, and 110k th, iterations for VOC 2007, VOC 2012, and MS-COCO, respectively. For comparison with other methods as per [86,85], we use data augmentation, including multi-level scaling and horizontal flipping during training. Multi-level scaling is also used during testing. In the augmentation, the shorter edges of input images undergo random re-scaling to dimensions selected from $\{480, 576, 688, 864, 1200\}$, while the longest edges are restricted to a maximum of 2,000. Additionally, a random horizontal flipping is applied to the scaled images. During evaluation, input images undergo augmentation using all five scales. The anchor generator is set to create anchors at sizes $\{32, 64, 128, 256, 512\}$ with aspect ratios of $\{0.5, 1.0, 2.0\}$. Moreover, we chose two fully connected layers to refine high-level features for precise object detection. Moreover, we standardized feature map sizes to 7×7 through RoI pooling, ensuring uniform processing regardless of initial object sizes.

In the CAM process, we use ResNet50 as our backbone and start by upsampling the maps to match the original image size and then apply a threshold of 0.5 to the map. The input image is normalized using mean values of [0.485, 0.456, 0.406] and standard deviations of $[0.229, 0.224, 0.225]$. To identify the segmented areas representing objects following post-thresholding, we leverage the

Table 1: SOTA WSOD vs. pseudo-labels bounding boxes from CAM-SAM ($mAP_{0.5}$ and mAP) on COCO dataset.

Methods	mAP	$mAP_{0.5}$
C-MIL [91]	8.5	19.4
WSOD2 [101]	10.8	22.7
C-MIDN [29]	9.6	21.4
MIST(+Reg+VGG16) [74]	11.4	24.3
MIST(+Reg+ResNet50) [74]	12.6	26.1
CASD +VGG16 [38]	12.8	26.4
CASD +ResNet50 [38]	13.9	27.8
Yin et al. [97]	13.6	27.6
CAM-SAM	**16.7**	**32.1**

Table 2: SOTA vs. CAM-SAM without train a detection model ($mAP_{0.5}$) on PASCAL VOC 2007 & 2012.

Methods	VOC 2007	VOC 2012
WSDDN [10]	34.8	-
OICR [86]	41.2	37.9
PCL [85]	43.5	40.6
C-MIL [91]	50.5	46.7
WSOD2(+Reg.) [101]	53.6	47.2
Pred Net [5]	52.9	48.4
C-MIDN [29]	52.6	50.2
MIST(+Reg.) [74]	54.9	52.1
CASD [38]	56.8	53.6
Yin et al. [97]	57.4	53.5
SLV [17]	53.5	49.2
IM-CFB [96]	54.3	49.4
CAM-SAM	**64.2**	**58.9**

`connectedComponentsWithStats` function from the OpenCV library, performing connected component analysis on an upsampled CAM with 8-connectivity. To create the point prompts for SAM, we adopt a sampling approach, evenly shifting by 15 pixels both vertically and horizontally from the pixel locations within the object area as provided by CAM after thresholding.

4.3 Quality of the Pseudo-Labels Bounding Boxes

The use of pseudo-labels comes with certain limitations, particularly regarding their quality and reliability. The limitations of pseudo-labels stem from the potential introduction of noise and inaccurate bounding boxes during their generation by SAM. Understanding the characteristics and potential biases of the SAM model producing the pseudo-labels is crucial for assessing their reliability and the subsequent impact on the performance of our detection model trained on these labels. Table 1 & 2 show the quality of the masks and the bounding boxes on different datasets comparing to SOTA in WSOD. The outcomes presented in Table 1 & 2 are based entirely on the use of our WSOD model in the first step, and these results were achieved without training an object detection model using pseudo-labels generated from CAM-SAM.

4.4 Quantitative Analysis of Noise Levels in CAM-SAM:

While, SAM model is highly accurate, yet the CAM-SAM approach for WSOD predominantly suffers from a high False Positive (FP) rate. By maintaining the sampling approach and evenly shifting by 15 pixels both vertically and horizontally, we document the FP rate in Table 3. In Table 3, CAM-SAM (w/o refinement) indicates the scenario where the pseudo-bounding boxes are used without applying our noise-refinement approach, whereas CAM-SAM (IDNR) refers to the case where we apply our instance-dependent noise refinement approach.

4.5 Study on the CAM-Variations

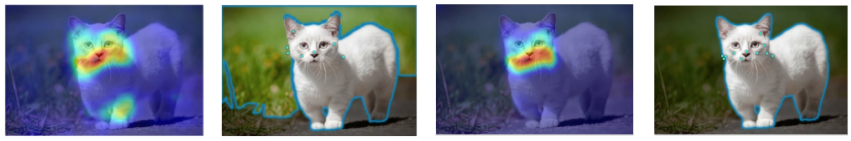
Exploring various adaptations of CAM [104,76,12], we specifically delved into three types the baseline CAM, Grad-CAM, and Grad-CAM++ [12]. Our ex-

Table 3: FP rate using $mAP_{0.5}$.

Methods	VOC 2007	VOC 2012	MS-COCO
CAM-SAM (w/o refinement)	17.7	29.9	40.7
CAM-SAM (IDNR)	12.1	18.8	32.1

Table 4: CAM Study on PASCAL VOC 2007 & 2012 and COCO with $mAP_{0.5}$.

Methods	VOC 2007	VOC 2012	MS-COCO
Uniform sampling-SAM	56.8	47.3	18.9
CAM-SAM (baseline)	61.1	55.9	29.8
CAM-SAM (GradCAM)	63.5	58.5	31.5
CAM-SAM (GradCAM++)	64.2	58.9	32.1



(a) (b)

Fig. 4: improved performance example with Grad-CAM++ over regular CAM.

periments show the superiority of Grad-CAM++ over the baseline CAM [104] and Grad-CAM [76] as reported in Table 4. This improvement potentially may be attributed to Grad-CAM++ leveraging a better pixel-wise contribution of gradients to activation, a contrast to the baseline CAM's reliance on global average pooling. To illustrate, we offer a perceptual example in Fig. 4 underscoring that Grad-CAM++ offers more precise attention in generating masks. We also tested uniform sampling (segment everything), which led SAM to over-segment the image. A classifier filtered out non-object areas, and the refined segments trained the detector. These experiments are performed without applying the instance-dependent noise refinement.

4.6 Multi-Labels Pre-trained vs. Per-Class Binary Classifiers

In this section, we evaluated our WSOD model using two different sources for generating the attention map for SAM. Firstly, we utilized a pre-trained ResNet50 classifier on ImageNet, selecting maps based on the top k classifier scores (usually up to 10 for COCO and 3 for PASCAL VOC) relative to the number of object classes in each training image. Note that in this experimental setting, we fine-tune the ResNet50 classifier on the aforementioned datasets. Secondly, we used a per-class binary classifier as detailed in section 3.1. This experiment was crucial for assessing the impact of attention map quality on SAM's performance. Our results, presented in Table 5, show that the per-class binary classifier setting, despite being more effort-intensive, significantly outperforms the pre-trained setting in terms of WSOD effectiveness. Note that these experiments are performed without applying the instance-dependent noise refinement strategy on the bounding boxes generated from the CAM-SAM step.

4.7 Training via Noisy and High-Confidence Bounding Boxes

In this section, we compare the performance of the detection model in cases where we train it with noisy labels versus high-confidence bounding boxes. Specifically,

Table 5: Pre-Trained Classifier (PTC) vs. Per-Class Classifier (PCC) with $mAP_{0.5}$.

Methods	VOC 2007	VOC 2012	MS-COCO
CAM-SAM (PTC)	63.5	57.1	29.8
CAM-SAM (PCC)	64.2	58.9	32.1

Table 6: Training via noisy and high-confidence bounding boxes with $mAP_{0.5}$.

Methods	VOC 2007	VOC 2012	MS-COCO
CAM-SAM (N)	66.3	60.9	33.8
CAM-SAM (HC)	67.9	62.1	34.5
CAM-SAM (IDNR)	71.1	64.4	36.1

Table 7: Comparison with the SOTA methods on PASCAL VOC 2012 for WSIS.

Methods	Backbone	$mAP_{0.25}$	$mAP_{0.5}$	$mAP_{0.7}$	$mAP_{0.75}$
Label-PEnet [30]	VGG-16	49.1	30.2	-	12.9
BESTIE [42]	HRNet48	61.2	51.0	31.9	26.6
PDSL [80]	ResNet50-WS	59.3	49.6	-	12.7
PRM [106]	ResNet-50	44.3	26.8	-	9.0
IAM [107]	ResNet-50	45.9	28.8	-	11.9
IRNet [3]	ResNet-50	-	46.7	23.5	-
WISE [46]	ResNet-50	49.2	41.7	-	23.7
Arun et al. [6]	ResNet-50	59.7	50.9	30.2	28.5
LIID [63]	ResNet-50	-	48.4	-	24.9
CAM-SAM (IDNR)	ResNet-50	**66.9**	**56.8**	**35.2**	**31.7**

after generating the noisy pseudo-label bounding boxes in the first step, we use them directly in training to progressively supervise our object detection model. While this approach is not perfect, it still shows improvement compared to the case where we solely annotate our images using CAM-SAM. The potential reason behind this improvement is the model's ability to progressively refine itself through self-training. In the second setting, we implemented a filtering strategy, discarding bounding boxes with a classification score below 0.8 as identified by a class-level trained classifier. Post filtering, we further fine-tuned the head of the Faster R-CNN with ResNet50 backbone. The compared results are indicated in Table 6. CAM-SAM (N) represents the scenario in which we train the detection model using noisy bounding boxes, while CAM-SAM (HC) denotes the scenario where we fine-tune the model with High-Confidence bounding boxes.

4.8 Contribution of the Instance-Dependent Noise Refinement

In this section, we explore the contribution of the second stage in our model where instance-dependent noise refinement is applied to the pseudo-labeled dataset generated in the first step. The training results, denoted as CAM-SAM (IDNR), are presented in Table 6. We utilized an AE architecture designed to learn sparse representations of the features obtained from the backbone. It comprises four key layers: an input layer, two hidden layers that compress the data, and an output layer tasked with reconstructing the data. Our objective is to minimize the discrepancy between the original input and its reconstruction, ensuring that the activations of the hidden layers remain sparse to highlight critical data features. To achieve this, our optimization framework employs a cost function that integrates reconstruction error and a sparsity penalty based on the activations of the hidden layer. This approach allows our AE to focus on the most significant aspects of the data. Here, we set the α, β, and p hyperparameters in Eq. 2 to 1, 0.001, and 0.01, respectively. Comparing the results in Table 6 indicates the effectiveness of noise refinement for training our detection model.

Table 8: Comparison with the SOTA methods on COCO dataset for WSIS.

Methods	Backbone	$mAP_{0.5}$	$mAP_{0.75}$	mAP
WS-JDS [81]	VGG16	11.7	5.5	6.1
JTSM [79]	ResNet18-WS	12.1	5.0	6.1
PDSL [80]	ResNet18-WS	13.1	5.0	6.3
CAM-SAM (Case 1)	ResNet-50	**33.4**	**16.6**	**17.1**
CAM-SAM (Case2)	ResNet-50	31.8	15.1	16.4
CAM-SAM (IDNR)	VGG16	22.6	9.4	10.3
CAM-SAM (IDNR)	ResNet18-WS	26.3	11.5	12.1
CAM-SAM (IDNR)	ResNet-50	29.9	12.8	14.7

Table 9: SSOD on COCO with mAP metric. All the results are the average of all 5 folds.

Methods	1%	5%	10%
STAC [82]	13.97 ± 0.35	24.38 ± 0.12	28.64 ± 0.21
ISMT [95]	18.88 ± 0.74	26.37 ± 0.24	30.53 ± 0.52
Instant Teaching [105]	18.05 ± 0.15	26.75 ± 0.05	30.40 ± 0.05
Unbiased Teacher [61]	20.75 ± 0.12	28.27 ± 0.11	31.50 ± 0.10
Soft Teacher [93]	20.46 ± 0.39	30.74 ± 0.08	34.04 ± 0.14
LabelMatch [13]	25.81 ± 0.28	32.70 ± 0.18	35.49 ± 0.17
CAM-SAM (IDNR)	**29.22 ± 0.15**	**36.15± 0.26**	**38.8 ± 0.32**

4.9 Weakly-Supervised Instance Segmentation Results

We initially used SAM for dataset annotation by generating segmentation masks, then adapted our model to instance segmentation by transitioning from Faster R-CNN to Mask R-CNN. Using our pseudo-labeling, we evaluated instance segmentation performance on COCO and PASCAL VOC 2012 and reported the results in Tables 7 & 8. Our CAM-SAM (IDNR) outperforms recent WSIS methods. Two ablation studies also indicate that using ground truth points (CAM-SAM-Case 1) or ground truth segmentation masks (CAM-SAM-Case 2) improve performance, but CAM-SAM (IDNR) still perform similarly well in WSIS.

4.10 Semi-Supervised Object Detection Results

We evaluate our model in a Semi-Supervised Object Detection (SSOD) scenario, combining labeled and unlabeled data to improve object detection. Our method uses a self-training technique [75] where the model assigns labels to unlabeled data and retrains iteratively. Initially, we annotate unlabeled images using our CAM-SAM (IDNR) method and apply multi-level scaling and horizontal flipping for data augmentation. Following the STAC protocol, we use 1%, 5%, and 10% of the training set as labeled data, with the rest as unlabeled. Results in Table 9 show the effectiveness of our model, CAM-SAM (IDNR), in generating high-quality pseudo-labels for the COCO dataset, outperforming competing methods.

5 Conclusion

We proposed a framework aimed at refining Weakly Supervised Object Detection (WSOD) using a pseudo-labeling strategy. Our method addresses label noise inherent in pseudo-label bounding boxes generated by the Segment Anything Model (SAM), which relies on cues from CAM. To mitigate this noise, we employ an instance-dependent, part-based noise correction model, enhancing our ability to discern accurate bounding boxes. Inspired by learning paradigms centered on part-based representations and human perception's tendency to simplify complex visuals, our method demonstrates effectiveness across various scenarios, including Semi-Supervised Object Detection (SSOD) and Weakly Supervised Instance Segmentation (WSIS), as evidenced by our experimental results.

References

1. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. IEEE transactions on pattern analysis and machine intelligence **26**(11), 1475–1490 (2004) 3, 6, 7, 8
2. Agarwal, S., Roth, D.: Learning a Sparse Representation for Object Detection. In: Proc. of the European Conference on Computer Vision (ECCV). pp. 113–128 (2002), http://cogcomp.org/papers/AgarwalRo02.pdf 3, 6, 7
3. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2209–2218 (2019) 4, 13
4. Algan, G., Ulusoy, I.: Image classification with deep learning in the presence of noisy labels: A survey. Knowledge-Based Systems **215**, 106771 (2021) 2
5. Arun, A., Jawahar, C., Kumar, M.P.: Dissimilarity coefficient based weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9432–9441 (2019) 1, 11
6. Arun, A., Jawahar, C., Kumar, M.P.: Weakly supervised instance segmentation by learning annotation consistent instances. In: European Conference on Computer Vision. pp. 254–270. Springer (2020) 4, 13
7. Ayinde, B.O., Zurada, J.M.: Deep learning of constrained autoencoders for enhanced understanding of data. IEEE transactions on neural networks and learning systems **29**(9), 3969–3979 (2017) 8
8. Berthon, A., Han, B., Niu, G., Liu, T., Sugiyama, M.: Confidence scores make instance-dependent label-noise learning possible. In: International conference on machine learning. pp. 825–836. PMLR (2021) 3
9. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological review **94**(2), 115 (1987) 3, 6, 8
10. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2846–2854 (2016) 1, 10, 11
11. Chadwick, S., Newman, P.: Training object detectors with noisy data. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 1319–1325. IEEE (2019) 4
12. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018) 11
13. Chen, B., Chen, W., Yang, S., Xuan, Y., Song, J., Xie, D., Pu, S., Song, M., Zhuang, Y.: Label matching semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14381–14390 (2022) 4, 14
14. Chen, L., Yang, T., Zhang, X., Zhang, W., Sun, J.: Points as queries: Weakly semi-supervised object detection by points. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8823–8832 (2021) 10
15. Chen, T., Mai, Z., Li, R., Chao, W.l.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.05803 (2023) 2
16. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. arXiv preprint arXiv:1904.04232 (2019) 10
17. Chen, Z., Fu, Z., Jiang, R., Chen, Y., Hua, X.S.: Slv: Spatial likelihood voting for weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12995–13004 (2020) 11

18. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1316–1326 (2023) 2
19. Cheng, J., Liu, T., Ramamohanarao, K., Tao, D.: Learning with bounded instance and label-dependent label noise. In: International conference on machine learning. pp. 1789–1799. PMLR (2020) 3, 8
20. Cheng, T., Wang, X., Chen, S., Zhang, Q., Liu, W.: Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3145–3154 (2023) 2, 4
21. Cholakkal, H., Sun, G., Khan, F.S., Shao, L.: Object counting and instance segmentation with image-level supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12397–12405 (2019) 4
22. Chun, D., Lee, S., Kim, H.: Usd: Uncertainty-based one-phase learning to enhance pseudo-label reliability for semi-supervised object detection. IEEE Transactions on Multimedia (2024) 2
23. Dai, Z., Cai, B., Lin, Y., Chen, J.: Up-detr: Unsupervised pre-training for object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1601–1610 (2021) 10
24. Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L.: Weakly supervised cascaded convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 914–922 (2017) 4
25. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997) 3
26. Ding, C.H., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE transactions on pattern analysis and machine intelligence **32**(1), 45–55 (2008) 7
27. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010) 10
28. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2009) 3, 6, 7
29. Gao, Y., Liu, B., Guo, N., Ye, X., Wan, F., You, H., Fan, D.: C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9834–9843 (2019) 4, 11
30. Ge, W., Guo, S., Huang, W., Scott, M.R.: Label-penet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3345–3354 (2019) 13
31. Ghosh, A., Kumar, H., Sastry, P.S.: Robust loss functions under label noise for deep neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017) 4
32. Guan, N., Liu, T., Zhang, Y., Tao, D., Davis, L.S.: Truncated cauchy non-negative matrix factorization. IEEE Transactions on pattern analysis and machine intelligence **41**(1), 246–259 (2017) 7

33. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018) 3, 4
34. He, C., Li, K., Zhang, Y., Xu, G., Tang, L., Zhang, Y., Guo, Z., Li, X.: Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. arXiv preprint arXiv:2305.11003 (2023) 2
35. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation **18**(7), 1527–1554 (2006) 8
36. Hosseini-Asl, E., Zurada, J.M., Nasraoui, O.: Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints. IEEE transactions on neural networks and learning systems **27**(12), 2486–2498 (2015) 3, 6, 8
37. Hsu, C.C., Hsu, K.J., Tsai, C.C., Lin, Y.Y., Chuang, Y.Y.: Weakly supervised instance segmentation using the bounding box tightness prior. Advances in Neural Information Processing Systems **32** (2019) 4
38. Huang, Z., Zou, Y., Kumar, B., Huang, D.: Comprehensive attention self-distillation for weakly-supervised object detection. Advances in neural information processing systems **33**, 16797–16807 (2020) 4, 11
39. Huang, Z., Bao, Y., Dong, B., Zhou, E., Zuo, W.: W2n: Switching from weak supervision to noisy supervision for object detection. In: European Conference on Computer Vision. pp. 708–724. Springer (2022) 4
40. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International conference on machine learning. pp. 2304–2313. PMLR (2018) 4
41. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 876–885 (2017) 4
42. Kim, B., Yoo, Y., Rhee, C.E., Kim, J.: Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4278–4287 (2022) 13
43. Kim, Y., Kim, J.M., Jeong, J., Schmid, C., Akata, Z., Lee, J.: Bridging the gap between model explanations in partially annotated multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3408–3417 (2023) 6
44. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 2
45. 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: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 1
46. Laradji, I.H., Vazquez, D., Schmidt, M.: Where are the masks: Instance segmentation with image-level supervision. arXiv preprint arXiv:1907.01430 (2019) 4, 13
47. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999) 7, 8
48. Lee, H., Ekanadham, C., Ng, A.: Sparse deep belief net model for visual area v2. Advances in neural information processing systems **20** (2007) 8

49. Lemme, A., Reinhart, R.F., Steil, J.J.: Online learning and generalization of parts-based image representations by non-negative sparse autoencoders. Neural Networks **33**, 194–203 (2012) 8
50. Li, G., Li, X., Wang, Y., Wu, Y., Liang, D., Zhang, S.: Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In: European Conference on Computer Vision. pp. 457–472. Springer (2022) 2
51. Li, J., Xiong, C., Socher, R., Hoi, S.: Towards noise-resistant object detection with noisy annotations. arXiv preprint arXiv:2003.01285 (2020) 4
52. Li, Q., Arnab, A., Torr, P.H.: Weakly-and semi-supervised panoptic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 102–118 (2018) 4
53. Li, X., Kan, M., Shan, S., Chen, X.: Weakly supervised object detection with segmentation collaboration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9735–9744 (2019) 3
54. Liao, M., Wan, F., Yao, Y., Han, Z., Zou, J., Wang, Y., Feng, B., Yuan, P., Ye, Q.: End-to-end weakly supervised object detection with sparse proposal evolution. In: European Conference on Computer Vision. pp. 210–226. Springer (2022) 1
55. Lin, J., Shen, Y., Wang, B., Lin, S., Li, K., Cao, L.: Weakly supervised open-vocabulary object detection. arXiv preprint arXiv:2312.12437 (2023) 2
56. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 10
57. Liu, C., Wang, K., Lu, H., Cao, Z., Zhang, Z.: Robust object detection with inaccurate bounding boxes. In: European Conference on Computer Vision. pp. 53–69. Springer (2022) 4
58. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023) 2
59. Liu, T., Gong, M., Tao, D.: Large-cone nonnegative matrix factorization. IEEE transactions on neural networks and learning systems **28**(9), 2129–2142 (2016) 7
60. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence **38**(3), 447–461 (2015) 9
61. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021) 4, 14
62. Liu, Y.C., Ma, C.Y., Kira, Z.: Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9819–9828 (2022) 4
63. Liu, Y., Wu, Y.H., Wen, P., Shi, Y., Qiu, Y., Cheng, M.M.: Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1415–1428 (2020) 4, 13
64. Logothetis, N.K., Sheinberg, D.L.: Visual object recognition. Annual review of neuroscience **19**(1), 577–621 (1996) 3, 6
65. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024) 2

66. Ma, X., Wang, Y., Houle, M.E., Zhou, S., Erfani, S., Xia, S., Wijewickrema, S., Bailey, J.: Dimensionality-driven learning with noisy labels. In: International Conference on Machine Learning. pp. 3355–3364. PMLR (2018) 4
67. Menon, A.K., Van Rooyen, B., Natarajan, N.: Learning from binary labels with instance-dependent noise. Machine Learning **107**, 1561–1595 (2018) 3
68. Nair, V., Hinton, G.E.: 3d object recognition with deep belief nets. Advances in neural information processing systems **22** (2009) 8
69. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. Advances in neural information processing systems **26** (2013) 3
70. Nguyen, T.D., Tran, T., Phung, D., Venkatesh, S.: Learning parts-based representations with nonnegative restricted boltzmann machine. In: Asian Conference on Machine Learning. pp. 133–148. PMLR (2013) 8
71. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: Proceedings of the IEEE international conference on computer vision. pp. 4930–4939 (2017) 1
72. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1944–1952 (2017) 3
73. Ranzato, M., Poultney, C., Chopra, S., Cun, Y.: Efficient learning of sparse representations with an energy-based model. Advances in neural information processing systems **19** (2006) 8
74. Ren, Z., Yu, Z., Yang, X., Liu, M.Y., Lee, Y.J., Schwing, A.G., Kautz, J.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10598–10607 (2020) 3, 11
75. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models (2005) 14
76. 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: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017) 11, 12
77. Seo, J., Bae, W., Sutherland, D.J., Noh, J., Kim, D.: Object discovery via contrastive learning for weakly supervised object detection. In: European Conference on Computer Vision. pp. 312–329. Springer (2022) 1
78. Shao, F., Chen, L., Shao, J., Ji, W., Xiao, S., Ye, L., Zhuang, Y., Xiao, J.: Deep learning for weakly-supervised object detection and localization: A survey. Neurocomputing **496**, 192–207 (2022) 2
79. Shen, Y., Cao, L., Chen, Z., Lian, F., Zhang, B., Su, C., Wu, Y., Huang, F., Ji, R.: Toward joint thing-and-stuff mining for weakly supervised panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16694–16705 (2021) 4, 14
80. Shen, Y., Cao, L., Chen, Z., Zhang, B., Su, C., Wu, Y., Huang, F., Ji, R.: Parallel detection-and-segmentation learning for weakly supervised instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8198–8208 (2021) 4, 13, 14
81. Shen, Y., Ji, R., Wang, Y., Wu, Y., Cao, L.: Cyclic guidance for weakly supervised joint detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 697–707 (2019) 14

82. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020) 4, 14
83. Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: International Conference on Machine Learning. pp. 5907–5915. PMLR (2019) 4
84. Tang, L., Xiao, H., Li, B.: Can sam segment anything? when sam meets camouflaged object detection. arXiv preprint arXiv:2304.04709 (2023) 2
85. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE transactions on pattern analysis and machine intelligence **42**(1), 176–191 (2018) 3, 10, 11
86. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2843–2851 (2017) 1, 3, 10, 11
87. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3132–3141 (2021) 4
88. Trigeorgis, G., Bousmalis, K., Zafeiriou, S., Schuller, B.: A deep semi-nmf model for learning hidden representations. In: International conference on machine learning. pp. 1692–1700. PMLR (2014) 7
89. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**, 154–171 (2013) 1
90. Wachsmuth, E., Oram, M., Perrett, D.: Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. Cerebral Cortex **4**(5), 509–522 (1994) 3, 6
91. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2199–2208 (2019) 11
92. Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., Sugiyama, M.: Part-dependent label noise: Towards instance-dependent label noise. Advances in Neural Information Processing Systems **33**, 7597–7610 (2020) 3, 7, 8
93. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3060–3069 (2021) 4, 14
94. Xu, Y., Zhu, L., Yang, Y., Wu, F.: Training robust object detectors from noisy category labels and imprecise bounding boxes. IEEE Transactions on Image Processing **30**, 5782–5792 (2021) 4
95. Yang, Q., Wei, X., Wang, B., Hua, X.S., Zhang, L.: Interactive self-training with mean teachers for semi-supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5941–5950 (2021) 4, 14
96. Yin, Y., Deng, J., Zhou, W., Li, H.: Instance mining with class feature banks for weakly supervised object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3190–3198 (2021) 1, 11
97. Yin, Y., Deng, J., Zhou, W., Li, L., Li, H.: Cyclic-bootstrap labeling for weakly supervised object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7008–7018 (2023) 1, 11

98. Yoo, J.H., Choi, S.J.: Nonnegative matrix factorization with orthogonality constraints. Journal of computing science and engineering **4**(2), 97–109 (2010) 7
99. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: International Conference on Machine Learning. pp. 7164–7173. PMLR (2019) 3
100. Yu, X., Liu, T., Gong, M., Tao, D.: Learning with biased complementary labels. In: Proceedings of the European conference on computer vision (ECCV). pp. 68–83 (2018) 3
101. Zeng, Z., Liu, B., Fu, J., Chao, H., Zhang, L.: Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8292–8300 (2019) 3, 11
102. Zhang, D., Liang, D., Yang, H., Zou, Z., Ye, X., Liu, Z., Bai, X.: Sam3d: Zero-shot 3d object detection via segment anything model. arXiv preprint arXiv:2306.02245 (2023) 2
103. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018) 4
104. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016) 2, 11, 12
105. Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-teaching: An end-to-end semi-supervised object detection framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4081–4090 (2021) 4, 14
106. Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J.: Weakly supervised instance segmentation using class peak response. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3791–3800 (2018) 4, 13
107. Zhu, Y., Zhou, Y., Xu, H., Ye, Q., Doermann, D., Jiao, J.: Learning instance activation maps for weakly supervised instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3116–3125 (2019) 4, 13
108. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 391–405. Springer (2014) 1