# Object Discovery via Contrastive Learning for Weakly Supervised Object Detection

Jinhwan Seo[1], Wonho Bae[2], Danica J. Sutherland[2,3], Junhyug Noh[4⋆], and Daijin Kim[1⋆]

[1] Pohang University of Science and Technology
[2] University of British Columbia
[3] Alberta Machine Intelligence Institute
[4] Lawrence Livermore National Laboratory
tohoaa@gmail.com, whbae@cs.ubc.ca, dsuth@cs.ubc.ca, noh1@llnl.gov, dkim@postech.ac.kr

**Abstract.** Weakly Supervised Object Detection (WSOD) is a task that detects objects in an image using a model trained only on image-level annotations. Current state-of-the-art models benefit from self-supervised instance-level supervision, but since weak supervision does not include count or location information, the most common "argmax" labeling method often ignores many instances of objects. To alleviate this issue, we propose a novel multiple instance labeling method called *object discovery*. We further introduce a new contrastive loss under weak supervision where no instance-level information is available for sampling, called *weakly supervised contrastive loss* (WSCL). WSCL aims to construct a credible similarity threshold for object discovery by leveraging consistent features for embedding vectors in the same class. As a result, we achieve new state-of-the-art results on MS-COCO 2014 and 2017 as well as PASCAL VOC 2012, and competitive results on PASCAL VOC 2007. The code is available at https://github.com/jinhseo/OD-WSCL.

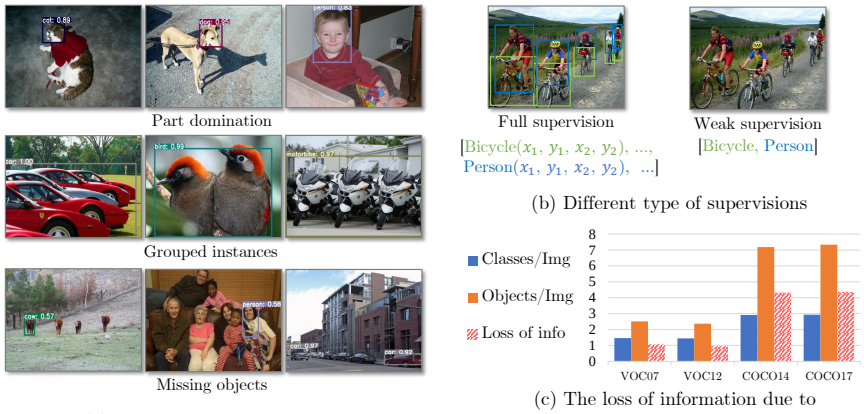
**Keywords:** Weakly Supervised Object Detection (WSOD)

## 1 Introduction

Object detection [21,19,20,18] has seen huge improvements since the introduction of deep neural networks and large-scale datasets [7,17,5]. It is, however, very expensive and time-consuming to annotate large datasets with fine-grained object bounding boxes. Recent work has thus attempted to use more cost-efficient annotations in an approach called Weakly Supervised Object Detection (WSOD), such as image, point, or "scribble" labels.

Although WSOD methods can be trained with much less annotation effort, however, the resulting models still perform far below their fully-supervised counterparts. We identify three categories of reasons for this deterioration, summarized in Fig. 1(a). First, *part domination* is when WSOD models focus only on

⋆ Corresponding authors: J. Noh (noh1@llnl.gov) and D. Kim (dkim@postech.ac.kr)




(a) Three challenges of WSOD

(b) Different type of supervisions

(c) The loss of information due to the argmax-based method

Fig. 1: (a) Three challenges of WSOD: part domination, grouped instances, and missing objects. (b) Unlike full supervision, there is no location and count information in weak supervision. (c) A large number of target objects are ignored in PASCAL VOC [7] and MS-COCO [17] due to the argmax-based method.

the discriminative part of an object, perhaps caused by the fundamentally ill-posed nature of framing WSOD as a Multiple Instance Learning (MIL) problem [6] prone to local minima, as done by much previous work [2,29,32]. The second major issue of WSOD is *grouped instances*, where neighbouring instances of objects in the same category are grouped into one large proposal, rather than proposed separately. As image-level annotations reveal only the presence of each object class, without any information about object location or counts (see Fig. 1(b)), it has become conventional to take only the single highest-score proposal as a "pseudo groundtruth" [2,29,12]. This can help avoid false positives, but often causes *missing objects*, where less-obvious instances are ignored.

Current argmax-based algorithms for finding pseudo groundtruths turn out to be problematic even on extremely popular benchmark datasets. Labeling only one proposal per category misses 40% of labels on PASCAL VOC [7] (selecting 7,306 of 12,608 target objects on VOC07), and 60% on MS-COCO [17] (533,396 of 894,204 objects on COCO14). Similar patterns hold for VOC12 and COCO17 (see Fig. 1(c)). Object detection models trained with this limited and, indeed, potentially confusing supervision are substantially hindered, and mining more pseudo-labels is problematic since they are likely to be false positives.

We introduce a novel multiple instance labeling method which addresses the limitations of current labeling methods in WSOD. Our proposed object discovery module explores all proposed candidates using a similarity measure to the highest-scoring representation. We further suggest a weakly supervised contrastive loss (WSCL) to set a reliable similarity threshold. WSCL encourages a model to learn similar features for objects in the same class, and to learn discriminative features for objects in different classes. To make sure the model

learn appropriate features, we provide a large number of positive and negative instances for WSCL through three feature augmentation methods suitable for WSOD. This well-behaved embedding space allows the object discovery module to find more reliable pseudo groundtruths. The resulting model then detects less-discriminative parts of target objects, misses fewer objects, and better distinguishes neighbouring object instances, as we will demonstrate experimentally in Section 5.1. As a result, the proposed approach beats state-of-the-art WSOD performance on both MS-COCO and PASCAL VOC by significant margins.

## 2 Related Work

### 2.1 Weakly Supervised Object Detection

Bilen *et al.* [2] introduced the first MIL-based end-to-end WSOD approach, known as WSDDN, which includes both classification and detection streams. Based on the MIL-based method, later works in WSOD have attempted to generate instance-level pseudo groundtruths in various ways.

**Self-Supervised Pseudo Labeling Approach.** To use instance-level supervision, Tang *et al.* [29] suggest Online Instance Classifier Refinement (OICR), which alternates between training instance classifier and selecting the most representative candidates. The online classifier rectifies initially detected instances through multiple stages and updates instance-level supervision determined by spatial relations. Tang *et al.* [28] expand clusters from OICR to include adjacent proposals that belong to the same cluster. Kosugi *et al.* [15] devise an instance labeling method to find positive instances based on a context classification loss, and to avoid negative instances using spatial constraints. Zeng *et al.* [36] show that the bottom-up evidence, unlike top-down class confidence, helps to recognize class-agnostic object boundaries. Chen *et al.* [4] propose a spatial likelihood voting (SLV) system to vote the bounding boxes with the highest likelihood in spatial dimension. Huang *et al.* [12] propose Comprehensive Attention Self-Distillation (CASD) that learns a balanced representation via input-wise and layer-wise feature learning. CASD aggregate attention maps, generated by multiple transformations and extracted from different levels of feature maps.

**Multiple Instance Approach.** Previous works focus on selecting valid pseudo groundtruths based on location information, but most still rely on the argmax labeling method which considers only one instance. Some attempts have been made, however, to provide multiple pseudo groundtruths. C-WSL [8] uses per-class object count annotations, which can help effectively separate grouped instances. OIM [16] exploits an object mining method to find undiscovered objects by calculating Euclidean distance between the core instance and its surrounding boxes. Ren *et al.* [22] propose not only Multiple Instance Self-Training (MIST) to generate top-k scored proposals as pseudo groundtruths, but also parametric dropblock to adversarially drop out discriminative parts. Yin *et al.* [35] introduce feature bank to provide one more pseudo groundtruth using the top-similarity scored instance. These algorithms [16,22,35] effectively find multiple instances per class, but their methods largely depend on heuristics, rather than learning.

Our proposed method instead explores all possible pseudo groundtruths in a more reliable way, with learning guided by a contrastive loss.

### 2.2 Contrastive Learning

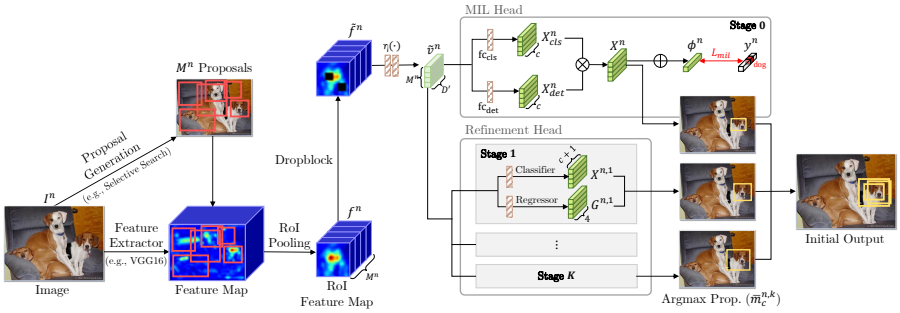
Contrastive losses have been successful in unsupervised and self-supervised learning for image classification tasks [3,14]. For object detection tasks, Xie *et al*. [33] and Sun *et al*. [27] demonstrate learning good embedding features via contrastive learning successfully improves the generalization ability of an object detector. One important factor of this success is to use effective mining strategies for positive and negative samples, which accelerates convergence and enhances the generalization ability of a model. In image classification tasks, these sample pairs are usually identified either by class labels [14] if available, or pairing images with versions that have been randomly altered with methods such as cropping, color distortion, or Gaussian blur [3]. Schroff *et al*. [24] introduce a hard positive and negative mining strategy based on the distance between anchor and positive samples, with full supervision. However, it is difficult to mine positive and negative samples in WSOD setting, which assumes no instance-level labels are available. Therefore, we propose feature augmentations to sample positives and negatives for contrastive learning in the WSOD setting. To the best of our knowledge, our method is the first approach to incorporate contrastive learning into WSOD tasks. Our proposed weakly supervised contrastive loss guides a model to learn consistent feature representations for objects in the same class and discriminative representations for ones in different classes, through mining positive and negative samples and augmenting intermediate features.

## 3 Background

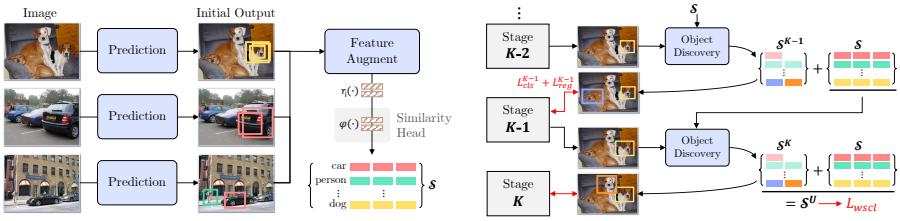
As with most state-of-the-art WSOD models, our approach is also based on the MIL head of WSDDN [2], followed by the refinement head suggested by OICR [29]. In this section, we describe how MIL and refinement heads work.

### 3.1 Feature Extractor

Let a batch $B = \{I^n, R^n, Y^n\}_{n=1}^N$ contain an image $I^n$, proposals $R^n = \{r_1, \dots, r_{M^n}\}$ with $M^n$ proposals for the image $I^n$, and image-level labels $Y^n = [y_1^n, ..., y_C^n] \in \{0,1\}^C$ where C is the number of classes. Given an image $I^n$, a feature extractor generates features for downstream tasks as follows. A backbone network takes a given image as input and outputs a feature map, from which a Region of Interest (RoI) feature map $f^n \in \mathbb{R}^{D \times H \times W \times M^n}$ is generated through an RoI pooling layer. Two fully-connected (FC) layers, which we denote $\eta(\cdot)$, map $f^n$ to RoI feature vectors $v^n \in \mathbb{R}^{D' \times M^n}$. To alleviate part domination, we also randomly mask out some blocks of the RoI map with Dropblock [22], generating $\tilde{f}^n$; we then generate regularized feature vectors $\tilde{v}^n = \eta(\tilde{f}^n)$. The MIL and refinement heads operate on $\tilde{v}^n$.




(a) Initial prediction step on single image ($I^n$)




(b) Sampling step for object discovery

(c) Object discovery step

Fig. 2: Overall architecture of the proposed method. Initial prediction in (a) collects top-scoring instances over all stages. Sampling step for Object Discovery in (b) iterates step (a) for all images in a batch, and applies feature augmentations described in Section 4.2. Object discovery in (c) mines additional pseudo groundtruths that are not recognized by the argmax method.

### 3.2 Multiple Instance Learning Head

As illustrated in Fig. 2(a), Multiple Instance Learning (MIL) head consists of classification and detection networks which take RoI feature vectors $\tilde{v}^n$ as input, and return classification scores $X^n_{cls} \in \mathbb{R}^{C\times M^n}$ and detection scores $X^n_{det} \in \mathbb{R}^{C\times M^n}$. Here, the $X^n_{cls}$ are computed by a softmax operation along the classes (rows), whereas $X^n_{det}$ are computed along the regions (columns). Proposal scores $X^n \in \mathbb{R}^{C\times M^n}$ are the element-wise product of classification and detection scores: $X^n = X^n_{cls} \odot X^n_{det}$. The image score of the $c$-th class, $\phi^n_c$, is obtained by the sum of proposal scores over all regions: $\phi^n_c = \sum_{m=1}^{M^n} X^n_{c,m}$. Given an image-level label $Y^n$ and image score $\phi^n_c$, the multi-label classification loss $L_{mil}$ is

$$L_{mil} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y^n_c \log \phi^n_c + (1-y^n_c)\log(1-\phi^n_c). \tag{1}$$

### 3.3 Refinement Head

The goal of the refinement head is to integrate self-supervised training strategy into WSOD via instance-level supervision. At the $k$-th stage ($k \in \{1, ..., K\}$),

an instance classifier generates proposal scores $X^{n,k} \in \mathbb{R}^{(C+1)\times M^n}$ where $M^n$ is the number of proposals and $C+1$ adds a background class to the $C$ classes. Instance-level supervision at the $k$-th stage is determined by previous stage; in particular, the first instance classifier takes supervision from the output of MIL head. Instance-level pseudo labels for the $n$-th image $Y^{n,k} \in \mathbb{R}^{(C+1)\times M^n}$ are then set to 1 if the corresponding proposal sufficiently overlaps the highest scored proposal, otherwise 0 as defined in (2).

$$\bar{m}_c^{n,k} = \underset{m}{\operatorname{argmax}}\, x_{c,m}^{n,(k-1)}; \qquad y_{c,m}^{n,k} = \begin{cases} 1 & \text{if } IoU(r_m, r_{\bar{m}_c^{n,k}}) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Finally, the instance classification loss $L_{cls}$ is defined as

$$L_{cls} = -\frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{M^n}\sum_{m=1}^{M^n}\sum_{c=1}^{C+1} w_m^{n,k} y_{c,m}^{n,k} \log x_{c,m}^{n,k} \tag{3}$$

where $x_{c,m}^{n,k}$ denotes $m$-th proposal score of a class $c$ at $k$-th stage, $w_m^{n,k}$ denotes a loss weight defined as $w_m^{n,k} = x_{c,\bar{m}_c^{n,k}}^{n,(k-1)}$ following OICR [29], and $K$ is the total number of refinement stages.

In addition to instance classification, some work [36,34] has improved localization performance by adding a bounding box regression loss. Given $\hat{M}^n$ pseudo groundtruth bounding boxes $\hat{g}_m^{n,k}$, nearby predicted bounding boxes $g_m^{n,k}$, matched as in (2), are encouraged to align using a $smooth_{L1}$ regression loss:

$$L_{reg} = -\frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\hat{M}^n}\sum_{m=1}^{\hat{M}^n} w_m^{n,k} smooth_{L1}(g_m^{n,k}, \hat{g}_m^{n,k}). \tag{4}$$

## 4 Our Approach

Most prior WSOD works [29,12] consider only top-scoring proposals, as described in (2). This strategy, however, has significant challenges in achieving our goal of detecting *all* objects present. To alleviate this issue, we propose a novel approach called *object discovery* which secures reliable pseudo groundtruths, by transferring instance-level supervision from previous to the next stage as illustrated in Section 3.3 and Fig. 2(c). To further enhance the object discovery module, we also introduce a new *similarity head* which maps RoI feature vectors to an embedding space, guided by a novel weakly supervised contrastive loss (WSCL).

### 4.1 Similarity Head

Parallel to the MIL and refinement heads, we construct a similarity head $\varphi(\cdot)$ that takes augmented RoI feature vectors as input described in Fig. 2. We will explain how RoI features are augmented in Section 4.2. The similarity head consists of two FC layers which map the inputs to a 128-dimensional space, followed by a normalization step. Thus, the outputs of the similarity head are expressed as $z^n = \varphi(v^n) \in \mathbb{R}^{128\times M^n}$ where $||z_m^n||_2 = 1$. Note that the similarity head uses $v^n$, whereas MIL and refinement head use the region-dropped $\tilde{v}^n$.

### 4.2 Sampling Strategy for Object Discovery

Contrastive learning [14,3], in general, focuses on making "positive" and "negative" pairs have similar and different feature embeddings, respectively. Although it is possible to augment images and pass each to the backbone to obtain pair of samples from RoI features in WSOD setting where no instance-level supervisions are available, it is computationally inefficient: most of the features from the backbone are not used as RoI features. Instead of augmenting images, we propose three feature augmentation methods to generate views of samples, as shown in Fig. 3(a): *IoU sampling*, *random masking*, and *adding gaussian noise*.

**IoU Sampling.** The purpose of IoU sampling is to increase the number of samples by treating the proposals adjacent to the top-scoring proposal $\bar{m}_c^n$ in (2) as positives. The proposals that overlap more than a threshold $\tau_{IoU}$ with the top-scoring proposal at each stage $k$ are considered positive samples, and the corresponding embedding vectors are formulated as

$$\begin{aligned} \mathcal{M}_c^{n,k} &= \{m \mid IoU(r_m, r_{\bar{m}_c^{n,k}}) > \tau_{IoU}, m = 1, 2, ..., M^n\} \\ \mathcal{Z}_{IoU}^{n,c} &= \{\varphi(\eta(f_m^n)) \mid m \in \bigcup_{k=0}^{K-1} \mathcal{M}_c^{n,k}\} \end{aligned} \tag{5}$$

where $M^n$ denotes the total number of proposals in $n$-th image, $\eta(\cdot)$ is the extractor of RoI feature vectors, and $\varphi(\cdot)$ denotes the similarity head.

**Random Masking.** Random masking randomly drops some regions across all channels of a RoI feature map. We first generate a random map $D : D_{i,j} \sim U(0,1) \in \mathbb{R}^{H \times W}$. Then the binary mask $D_{drop}$ is determined by drop threshold $\tau_{drop}$, so if $D < \tau_{drop}$, $D_{drop}$ is set to be 0, otherwise 1. Finally, a randomly-masked feature is obtained by taking spatial-wise multiplication of a RoI feature map $f_m^n$ with $D_{drop}$ followed by the similarity head,

$$\mathcal{Z}_{mask}^{n,c} = \{\varphi(\eta(f_m^n \odot D_{drop})) \mid m \in \bigcup_{k=0}^{K-1} \mathcal{M}_c^{n,k}\}. \tag{6}$$

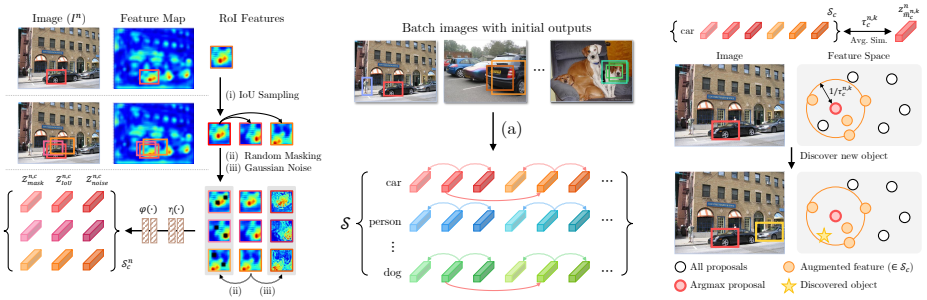
Here, random masking is applied to the all proposals from IoU sampling at all stages, $\bigcup_{k=1}^{K} \mathcal{M}_c^{n,k}$, to obtain more positive samples.

**Adding Gaussian Noise.** To add Gaussian random noise to RoI feature maps, we create a random noise map $D_{noise} : D_{i,j} \sim N(0,1) \in \mathbb{R}^{H \times W}$. We add this to the RoI feature maps $f_m^n$ by

$$\mathcal{Z}_{noise}^{n,c} = \{\varphi(\eta(f_m^n + f_m^n \odot D_{noise})) \mid m \in \bigcup_{k=0}^{K-1} \mathcal{M}_c^{n,k}\}. \tag{7}$$

**Cross-Image Representations.** Finally, we gather the augmented embedding vectors corresponding to the same object category from different images, as described in Fig. 3(b). We treat cross-batch representations from the same categories as positive examples in the mini-batch,

$$\mathcal{S}_c = \bigcup_{n=1}^{N} (\mathcal{Z}_{IoU}^{n,c} \cup \mathcal{Z}_{mask}^{n,c} \cup \mathcal{Z}_{noise}^{n,c}). \tag{8}$$




(a) Feature augmentation (b) Cross-image representations (c) Object discovery

Fig. 3: Three steps for object discovery module. (a) applies feature augmentation methods to embedding vectors of top-scoring proposals. (b) collects all augmented embedding vectors through all images in a batch. (c) determines new pseudo groundtruths based on the similarity with the embedding vector of the top-scoring instance $z^n_{\bar{m}_c^{n,k}}$ and similarity threshold $\tau_c^{n,k}$.

### 4.3 Object Discovery

Using the augmented RoI features introduced in the previous section, we discover many reliable instance-level pseudo groundtruths missed by previous methods that take only the top-scoring proposals. Intuitively, even though the classification score of a proposal may be low, if its embedding vector is close to that of the top-scoring proposal, it is likely that the proposal shares the same class as the top-scoring proposal. Therefore, instead of solely relying on classification scores, we exploit similarity scores between the embedding vectors of all proposals and the top-scoring (argmax) proposal at each stage $k$, to discover additional pseudo grountruths as shown in Fig. 3(c). To mine new pseudo groundtruths, we first compute a threshold $\tau_c^{n,k}$ that determines whether to label a proposal as a pseudo groundtruth for a class $c$ at stage $k$:

$$\tau_c^{n,k} = \frac{1}{|\mathcal{S}_c|} \sum_{i=1}^{|\mathcal{S}_c|} sim(z^n_{\bar{m}_c^{n,k}}, \mathcal{S}_{c,i}), \tag{9}$$

where $z^n_{\bar{m}_c^{n,k}}$ denotes embedding vectors of top-scoring proposals at stage $k$, $S_{c,i}$ denotes $i$-th element of $S_c$, and $sim(\cdot,\cdot)$ is a dot product between inputs. Then, new pseudo groundtruth candidates $\acute{\mathcal{M}}_c^{n,k}$ are determined as the ones having higher similarity to the top-scoring proposal than similarity threshold $\tau_c^{n,k}$,

$$\acute{\mathcal{M}}_c^{n,k} = \{m \mid sim(z^n_m, z^n_{\bar{m}_c^{n,k}}) > \tau_c^{n,k}, m = 1, 2, ..., M^n\}. \tag{10}$$

Finally, we add new pseudo groundtruths denoted as $\tilde{\mathcal{M}}_c^{n,k}$ at stage k after applying Non-Maximum Suppression (NMS) [23] to $\acute{\mathcal{M}}_c^{n,k}$.

Consequently, we update instance-level supervision and embedding vectors for newly discovered pseudo groundtruths $\tilde{\mathcal{M}}_c^{n,k}$. We re-label instance-level supervision $\{y_{c,m}^{n,k} | m \in \tilde{\mathcal{M}}_c^{n,k}\}$ and its adjacent proposals as described in (2). New

embedding vectors $\{z_m^n | m \in \tilde{\mathcal{M}}_c^{n,k}\}$ are add to $S_c^k$, as it is expected discriminative features help to measure precise similarity. Then, classification loss $L_{cls}$ in (3) and regression loss $L_{reg}$ in (4) are updated accordingly. After going through all the stages from 1 to $K$, we obtain $S^U = S^K \cup S$ as described in Fig. 2(c).

### 4.4 Weakly Supervised Contrastive Loss

To learn more consistent feature representations for the proposals in the same class, we propose weakly supervised contrastive loss (WSCL) that learns representations by attracting positive samples closer together and repelling negative samples away from positives samples in the embedding space. From a collection $S^U = \{s_i, t_i\}_{i=1}^{|S^U|}$ where $s_i$ denotes $i$-th embedding vectors, $t_i$ denotes the pseudo label of $s_i$, WSCL for $i$-th embedding vector denoted as $L_{wscl}^i$, is formulated as

$$L_{wscl}^i = -\frac{1}{N_{t_i} - 1} \sum_{j=1, j \neq i}^{|S^U|} \mathbb{1}\{t_i = t_j\} \cdot \log \frac{\exp\left(s_i \cdot s_j / \epsilon\right)}{\sum_{l=1, l \neq i}^{|S^U|} \exp\left(s_i \cdot s_l / \epsilon\right)} \tag{11}$$

where $N_{t_i} \coloneqq \sum_{j=1}^{|S^U|} \mathbb{1}\{t_i = t_j\}$, and $\epsilon$ is a temperature parameter introduced in [14]. Note that $\mathcal{S}^U = \bigcup_{c=1}^{C} \mathcal{S}_c^U$.

**Instance difficulty.** Since confidence score of instance is noisy at early stages of training, we introduce instance difficulty $\omega$ to make training for WSCL easier. $\omega$ is the set of scores for all images in a batch where each score is the instance score from the MIL head over the sum of them at each image. Here, the size of $\omega$ is the same as $\mathcal{S}$. Then, the re-weighted contrastive loss is formulated as,

$$L_{wscl} = \frac{1}{|S^U|} \sum_{i=1}^{|S^U|} \omega_i \cdot L_{wscl}^i, \; \omega = \bigcup_{n=1}^{N} \left\{ X_{c,m}^n / \sum_{j=1}^{M^n} X_{c,j}^n \mid m \in \bigcup_{k=0}^{K-1} \mathcal{M}_c^{n,k} \right\} \tag{12}$$

**Total Loss.** Finally, the total loss of training the proposed model is defined as

$$L_{total} = L_{mil} + L_{cls} + L_{reg} + \lambda L_{wscl} \tag{13}$$

where $\lambda$ is a loss weight to balance scale with the other losses.

## 5 Experiments

### 5.1 Experiment Setting

**Datasets.** To verify the robustness of our method, we evaluate it on four object detection datasets; VOC07 and VOC12 in PASCAL VOC [7], and COCO14 and COCO17 in MS-COCO [17], following the convention in WSOD tasks. We use trainval sets containing 5,011 and 11,540 images for VOC07 and VOC12, respectively, and test sets that contain 4,951 and 10,991 images, for evaluation.

We further investigate the robustness of our method on MS-COCO datasets. For COCO14, we train our model on the train set of 82,783 images and test it with the validation set of 40,504 images. For COCO17, we split the dataset into the train set of 118,287 images and validation set of 5,000 images. We use only image-level annotations to train our model on all datasets.

**Evaluation Metrics.** On VOC07 and 12 datasets, we evaluate our model on the test set using mean Average Precision (mAP) metric with standard IoU criterion (0.5). MS-COCO is more challenging than PASCAL VOC as it has significantly more instances per image (about 2 *vs*. 7) and more classes (20 *vs*. 80). For this reason, MS-COCO is often not considered in the WSOD literature. We report the performance on MS-COCO datasets following the standard COCO metric which includes several metrics, such as, average precision (AP) and average recall (AR) with varying IoU thresholds *e.g*., 0.5 and 0.75, and object sizes *e.g*., small ($s$), medium ($m$), and large ($l$), but the most representative metric is the AP averaged over 10 IoU thresholds (from 0.5 to 0.95 for every 0.05 step).

**Implementation Details.** All the implementation is developed in PyTorch. For both VGG16 [26] and ResNet [11] models, we initialize parameters using ImageNet [5] pre-trained networks. For VGG16, following the previous methods [22,12], we replace a global average pooling layer with a RoI pooling layer, and remove the last FC layer leaving two FC layers, which all the heads including the similarity head are attached to. For ResNet, we modify the structure for WSOD as suggested in Section 4 of Shen *et al*. [25]. We use around 2,000 proposals per image for both proposal generation methods, SS [31] and MCG [1].

**Hyperparameters.** The batch size is set to 8 for PASCAL VOC and MS-COCO datasets. We train models for 30K, 60K and 130K iterations on VOC07, VOC12 and MS-COCO, respectively, using a SGD optimizer with the learning rate of 0.01 and weight decay of 0.0001 except for R50-WS and R101-WS [25] where the learning rate is set to 0.02 on both datasets. At inference time, the prediction scores are computed as the average of proposal scores for all $k$-stages, and the offsets from regression branch are incorporated to adjust the coordinates of bounding boxes. The final predictions are made after applying NMS of which threshold is set to 0.4 for both datasets. Following the previous methods [22,29,12], the inputs are multi-scaled to $\{480, 576, 688, 864, 1000, 1200\}$ for both training and inference time. In the experiments, we set $\tau_{IoU} = 0.5$, $\tau_{drop} = 0.3$, $\tau_{nms} = 0.1$, $\lambda = 0.03$ ($\lambda = 0.01$ on COCO datasets) and $\epsilon = 0.2$ for WSCL and K=3 for the number of refinement stages. But, as we will show in an ablation study, the performance is not sensitive to the choice of hyperparameters.

### 5.2 Quantitative Results

**Comparison with state-of-the-arts.** In Table 1, we compare the proposed method with other state-of-the-art algorithms on COCO14 and 17. Regardless of backbone structure and dataset, our method achieves the new state-of-the-art performance for all the evaluation metrics. The fact that the performance of Ours in AR measurements are higher than the other methods implies that the proposed method successfully detects missing instances compared to the previous

Table 1: Comparison of the state-of-the-art algorithms on MS-COCO

| Dataset | Backbone | Method | $AP$ | $AP^{50}$ | $AP^{75}$ | $AP^s$ | $AP^m$ | $AP^l$ | $AR^1$ | $AR^{10}$ | $AR^{100}$ | $AR^s$ | $AR^m$ | $AR^l$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| COCO14 | VGG16 | PCL [28] | 8.5 | 19.4 | - | - | - | - | - | - | - | - | - | - |
| | | C-MIDN [9] | 9.6 | 21.4 | - | - | - | - | - | - | - | - | - | - |
| | | WSOD2 [36] | 10.8 | 22.7 | - | - | - | - | - | - | - | - | - | - |
| | | MIST [22] | 11.4 | 24.3 | 9.4 | 3.6 | 12.2 | 17.6 | 13.5 | 22.6 | 23.9 | 8.5 | 25.4 | 38.3 |
| | | CASD [12] | 12.8 | 26.4 | - | - | - | - | - | - | - | - | - | - |
| | | Ours | **13.7** | **27.7** | **11.9** | **4.4** | **14.5** | **21.2** | **14.7** | **24.8** | **26.9** | **8.8** | **27.8** | **44.0** |
| | ResNet50 | MIST [22] | 12.6 | 26.1 | 10.8 | 3.7 | 13.3 | 19.9 | 14.8 | 23.7 | 24.7 | 8.4 | 25.1 | 41.8 |
| | | CASD [12] | **13.9** | 27.8 | - | - | - | - | - | - | - | - | - | - |
| | | Ours | **13.9** | **29.1** | **11.8** | **4.9** | **16.8** | **22.3** | **15.5** | **26.1** | **28.0** | **9.0** | **31.8** | **46.6** |
| | ResNet101 | MIST [22] | 13.0 | 26.1 | 10.8 | 3.7 | 13.3 | 19.9 | 14.8 | 23.7 | 24.7 | 8.4 | 25.1 | 41.8 |
| | | Ours | **14.4** | **29.0** | **12.4** | **4.8** | **17.3** | **23.8** | **15.8** | **27.0** | **30.0** | **9.2** | **33.6** | **51.0** |
| COCO17 | VGG16 | MIST [22] | 12.4 | 25.8 | 10.5 | 3.9 | 13.8 | 19.9 | 14.3 | 23.3 | 24.6 | **9.7** | 26.6 | 39.6 |
| | | Ours | **13.6** | **27.4** | **12.2** | **4.9** | **15.5** | **21.6** | **14.6** | **24.8** | **26.8** | 9.2 | **28.7** | **43.8** |
| | ResNet50 | Ours | 13.8 | 27.8 | 12.1 | 5.7 | 17.7 | 23.8 | 15.1 | 26.6 | 29.7 | 10.1 | 33.7 | 50.7 |
| | ResNet101 | Ours | 14.4 | 28.7 | 12.6 | 5.4 | 17.9 | 25.5 | 15.4 | 26.8 | 29.6 | 10.0 | 33.3 | 50.6 |

methods on MS-COCO, which contain many instances and categories per image. For instance, our method outperforms MIST [22] by a large margin in AR measurements; on average of 2.13% for $(AR^1, AR^{10}, AR^{100})$ and on average of 3.46% for $(AR^s, AR^m, AR^l)$ with VGG16 on COCO14. Despite the significant improvement in AR, AP also improves by a large margin regardless of different subcategory of AP and backbone structure. Our method gains on average of 2.95% for $(AP^{50}, AP^{75})$ and on average of 2.23% for $(AP^s, AP^m, AP^l)$ with VGG16 on COCO14. We observe similar tendency on COCO17 for all the backbones. As a result, our proposed method achieves the new state-of-the-art performance on both COCO14 and COCO17 (14.4%).

In Table 2, we compare the performance of the state-of-the-art methods on PASCAL VOC with both SS and MCG proposal methods. Since PASCAL VOC datasets contain less number of instances and categores per image, the gain achieved by our method is relatively lower than MS-COCO. It, however, outperforms other multiple instance labeling methods [8,16,22] with clear margins. For example, Ours outperforms MIST [22] (the best performance multiple instance labeling method) by 1.2% and 2.5% on VOC07 and VOC12, respectively, with SS proposal method, and 2.2% and 2.3% with MCG proposal method. It also achieves the new state-of-the-art performance on VOC12 (54.6%) and compatible results on VOC07 (Ours: 56.1% *vs*. CASD: 56.8%). We use MCG for the rest of the experiments as it outperforms SS.

Table 2: Comparison of the state-of-the-art methods on PASCAL VOC.

| Proposal | Method | VOC07 | VOC12 |
|---|---|---|---|
| SS[31] | WSDDN[2] | 34.8 | - |
| | OICR[29] | 41.2 | 37.9 |
| | PCL[28] | 43.5 | 40.6 |
| | C-WSL[8] | 46.8 | 43.0 |
| | WSRPN[30] | 47.9 | 43.4 |
| | C-MIL[32] | 50.5 | 46.7 |
| | C-MIDN[9] | 52.6 | 50.2 |
| | WSOD2[36] | 53.6 | 47.2 |
| | OIM[16] | 50.1 | 45.3 |
| | SLV[4] | 53.5 | 49.2 |
| | MIST[22] | 54.9 | 52.1 |
| | CASD[13] | **56.8** | 53.6 |
| | Ours | 56.1 | **54.6** |
| MCG[1] | MIST[22] | 56.5 | 53.9 |
| | CASD[13] | 57.4 | - |
| | Ours | **58.7** | **56.2** |

Table 3: Experiment results with various settings on VOC07. (a) Different components of the proposed method. (b) Performance with ResNet backbones [25]. (c) Comparison of different combination of feature augmentation methods.

| OD | WSCL | mAP |
|---|---|---|
| | | 52.3 |
| ✓ | | 56.1 |
| | ✓ | 54.5 |
| ✓ | ✓ | **58.7** |

(a) Diff. components

| Method | R50-WS | R101-WS |
|---|---|---|
| OICR[29] | 50.9 | 51.4 |
| PCL[28] | 50.8 | 53.3 |
| C-MIL[32] | 53.4 | 53.9 |
| Ours | **56.6** | **56.5** |

(b) ResNet backbones

| Method | IoU | Mask | Noise | mAP |
|---|---|---|---|---|
| OICR$^+$ | | | | 52.3 |
| | ✓ | | | 56.8 |
| | | ✓ | | 55.1 |
| | | | ✓ | 54.8 |
| + Ours | ✓ | ✓ | | 57.4 |
| | ✓ | | ✓ | 57.4 |
| | | ✓ | ✓ | 55.2 |
| | ✓ | ✓ | ✓ | **58.7** |

(c) Feature augmentation

**The effectiveness of each component.** To validate the effectiveness of each component of Object Discovery (OD) and WSCL modules, we provide experiment results on VOC07 with SS in Table 3(a). Note that we find the performance of OICR [29] can be further increased by adding the bounding box regression and dropblock [10] layers (45.0% $\rightarrow$ 52.3% on VOC07), thus, we call OICR + Regression + Dropblock as OICR$^+$, and use it as the baseline throughout the experiments unless specified otherwise. In Table 3(a), each of OD and WSCL modules significantly improves OICR$^+$ baseline (+3.8 and +2.2) but the improvement is the highest when both OD and WSCL are applied simultaneously, which partially demonstrate that each module helps the other one as described in Section 4.

**Robust regardless of backbone.** In Table 1, we have shown that the proposed method performs well both in AP and AR metrics on MS-COCO datasets regardless of backbone structure. Although it is not a common practice to provide the performance with ResNet backbones on PASCAL VOC, we provide the performance of our method on ResNet in Table 3(b). It shows that Ours with ResNet backbones significantly outperforms the previous state-of-the-art methods as with the case with VGG backbone shown in Table 2. The performance of the previous methods are reported in [25].

### 5.3 Ablation Studies

**Feature augmentation methods.** To further verify the effectiveness of the proposed feature augmentation methods, we experiment with different combination of augmentation methods for the object discovery and WSCL modules in Table 3(c). The performance significantly improves from 52.3% to 55.1% and 54.8% with random masking and Gaussian noise, respectively, even without IoU sampling. With IoU sampling prior to random masking and Gaussian noise, the performance consistently improves further by 2.3% and 2.6% for random masking and Gaussian noise. Using all the proposed feature augmentations, the performance reaches 58.7% that is 6.4% higher than the baseline.

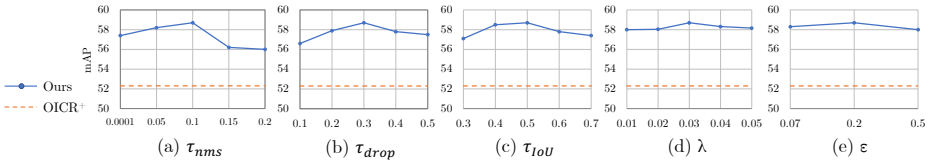
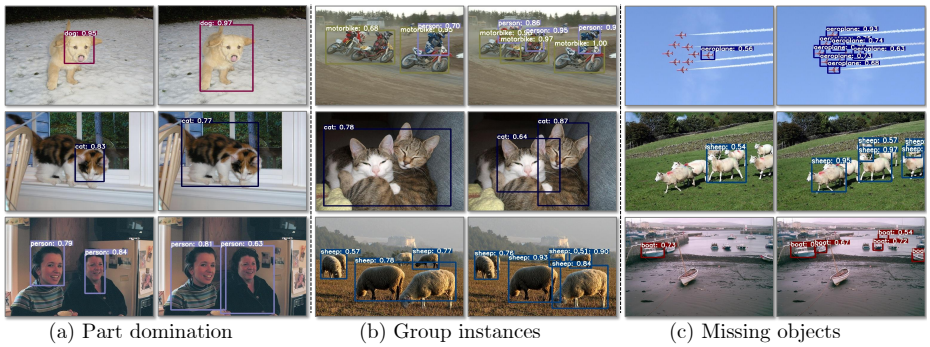



Fig. 4: Performance with different values of $\tau_{nms}$, $\tau_{drop}$, $\tau_{IoU}$, $\lambda$ and $\epsilon$.



(a) Part domination (b) Group instances (c) Missing objects

Fig. 5: Qualitative results of OICR [29] and ours about the three challenges of WSOD: (a) part domination, (b) grouped instances and (c) missing objects. The images on the left and right indicate OICR and Ours, respectively.

**Sensitivity to hyperparameters.** In Fig. 4, we provide the experiment results with different values of the hyperparameters we introduce. Regardless of hyperparameter, the performance is not sensitive to the choice of values around the optimal values we choose ($\tau_{nms} = 0.1, \tau_{drop} = 0.3, \tau_{IoU} = 0.5, \lambda = 0.03$ and $\epsilon = 0.2$). For instance, the gap between the highest and lowest performance for each hyperparameter is no more than 2.6% in mAP (highest with $\tau_{nms}$), which demonstrates that our proposed method does not greatly depend on hyperparameter tuning. In (e), we use the same values of $\epsilon$ following the experiments conducted in other contrastive learning methods [14,3].

### 5.4 Qualitative results

**Three challenges of WSOD.** In Fig. 5, we provide the qualitative results that show how our method addresses three main challenges of WSOD – part domination, grouped instances and missing objects (described in Section 1), compared to OICR [29]. The left columns show the results from OICR [29] whereas the right columns show the results from our method. The part domination shown in (a) is largely alleviated, especially for the categories with various poses such as *dog*, *cat* and *person*. We also observe that grouped instances are separated into multiple bounding boxes in (b). Lastly, our method successfully detects many of instances that are ignored with the argmax labeling method as shown in (c).

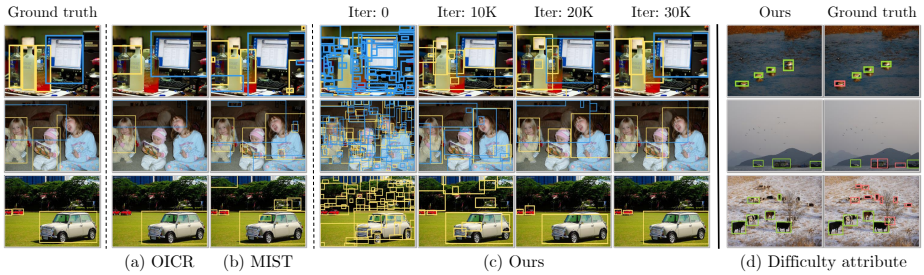



Fig. 6: Comparison of pseudo groundtruths generated by (a) OICR [29], (b) MIST [22] and (c) Ours which shows pseudo groundtruths at different training steps. "Difficult" objects are often captured by Ours as shown in (d).

**Selection of pseudo groundtruths.** We visualize the pseudo groundtruths captured by OICR [29], MIST [22] and Ours in Fig. 6. OICR [29] selects only the top-scoring proposal per category ignoring all the other instances as shown in (a). Although multiple objects are captured by MIST [22] in (b), it also selects many false positives *e.g.*, object-like background. Ours also captures many false positives in early stages of training (Iter: 0 – 10K) but later in training, it mostly selects true positives (Iter: 20K – 30K). Our method can even detect some objects categorized as "difficult" (red boxes in (e)), which are not considered for detection performance as they are too hard even for humans to detect.

## 6 Conclusion

We propose a novel multiple instance labeling method to replace the conventional argmax-based pseudo groundtruth labeling method for weakly supervised object detection (WSOD). To this end, we introduce a contrastive loss for the WSOD setting that learns consistent embedding features for proposals in the same class, and discriminative features for ones in different classes. With these features, it is possible to mine a large number of reliable pseudo groundtruths, which provide richer supervision for WSOD tasks. As a result, we achieve the new state-of-the-art results on both PASCAL VOC and MS-COCO benchmarks.

**Acknowledgements**. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.2017-0-00897, Development of Object Detection and Recognition for Intelligent Vehicles) and (No.B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis), as well as by support provided by the Natural Sciences and Engineering Research Council of Canada and the Canada CIFAR AI Chairs program. Junhyug Noh was supported by LLNL under Contract DE-AC52-07NA27344.

## References


1. Arbeláez, P., Pont-Tuset, J., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 328–335 (2014)
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. 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)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. 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)
7. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (Jan 2015)
8. Gao, M., Li, A., Yu, R., Morariu, V.I., Davis, L.S.: C-wsl: Count-guided weakly supervised localization. In: The European Conference on Computer Vision (ECCV) (September 2018)
9. 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)
10. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. arXiv preprint arXiv:1810.12890 (2018)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Huang, Z., Zou, Y., Bhagavatula, V., Huang, D.: Comprehensive attention self-distillation for weakly-supervised object detection. arXiv preprint arXiv:2010.12023 (2020)
13. Hwang, J., Kim, S., Son, J., Han, B.: Weakly supervised instance segmentation by deep community learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1020–1029 (2021)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
15. Kosugi, S., Yamasaki, T., Aizawa, K.: Object-aware instance labeling for weakly supervised object detection. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
16. Lin, C., Wang, S., Xu, D., Lu, Y., Zhang, W.: Object instance mining for weakly supervised object detection. In: AAAI. pp. 11482–11489 (2020)
17. 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: European conference on computer vision. pp. 740–755. Springer (2014)

18. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
20. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
22. 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)
23. Rosenfeld, A., Thurston, M.: Edge and curve detection for visual scene analysis. IEEE Transactions on computers **100**(5), 562–569 (1971)
24. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
25. Shen, Y., Ji, R., Wang, Y., Chen, Z., Zheng, F., Huang, F., Wu, Y.: Enabling deep residual networks for weakly supervised object detection. In: European Conference on Computer Vision. pp. 118–136. Springer (2020)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
27. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7352–7362 (2021)
28. 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)
29. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
30. Tang, P., Wang, X., Wang, A., Yan, Y., Liu, W., Huang, J., Yuille, A.: Weakly supervised region proposal network and object detection. In: The European Conference on Computer Vision (ECCV) (September 2018)
31. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2), 154–171 (2013)
32. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
33. Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., Li, Z., Luo, P.: Detco: Unsupervised contrastive learning for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8392–8401 (2021)
34. Yang, K., Li, D., Dou, Y.: Towards precise end-to-end weakly supervised object detection network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

35. 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. pp. 3190–3198 (2021)
36. 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)