Transformer Based Weakly Supervised Framework for Multi-Scaled Object Detection

Shinhye Han[1] Jeongwoo Shin[1] Keunyoung Kim[1]
[1]Seoul National University, Republic of Korea
{sienna.shhan, swswss, keunyoung.kim}@snu.ac.kr

Abstract

Weakly supervised object detection (WSOD) has gained great attention as a remarkable topic in the field of computer vision in the aspect of data-efficient method to avoid expensive hand-crafted annotations. However, due to the lack of bounding box information, state-of-the-art deep neural networks still suffer from detecting multi-scaled objects, especially small ones. To mitigate this problem, our study propose attention and similarity based pseudo bounding box generator to collectively detect multi-scaled objects, leading the model to employ class-agnostic representations of objects. Armed with this novel approach, Transformer based DETR architecture which deploys multi-instance head and refinement head, present plausible result on COCO small object detection task and favorable results on COCO 2014, and PASCAL VOC 2007. Experimental results demonstrate the effectiveness of our proposed methods in detecting multi-scaled objects in WSOD manner.

1. Introduction

Object detection has achieved impressive improvements with the advent of deep neural networks and large datasets [30, 33, 34]. However, detecting multi-scaled objects, especially small objects remains a challenge in object detection despite of their various applications, such as large-scale monitoring or surveillance [16, 53], and assistance for autonomous driving [4, 61] or diagnosis [19]. Small objects have fewer visible features since they consist of lesser pixels (under 32 pixels), which can lead to difficulties in detecting them [11, 18]. Accordingly, the primary objective of this study is to detect both small and large objects collectively.

To overcome the challenges of detecting multi-scaled objects, deep learning-based models require a large amount of annotations. However, obtaining instance-level annotations can be highly time-consuming and labor-intensive. Therefore, weakly supervised object detection (WSOD) has emerged as an alternative method, where models use weaker supervision such as image-level labels to produce bounding box-level predictions [24,42]. This method can alleviate the necessity of hard supervision for object detection.

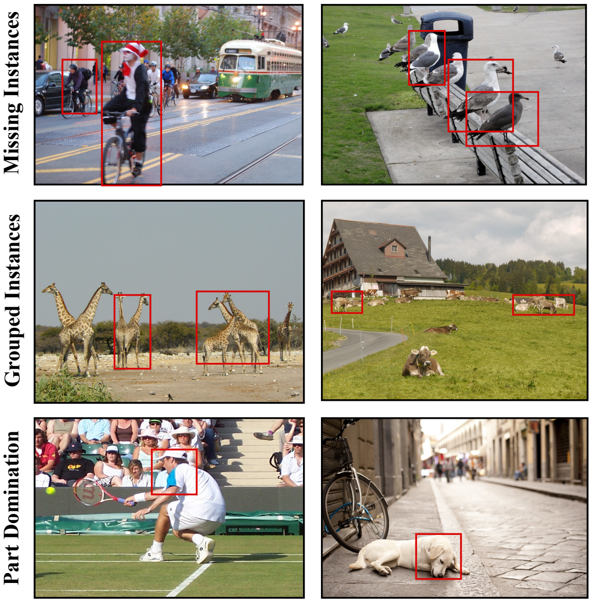

Figure 1. **Three issues of WSOD:** (1) missing instances (top), (2) grouped instances (middle), (3) part domination (bottom)

Despite its usefulness, it is usually accompanied by three adversarial issues: missing instances, grouped instances, and part domination [35,39] as illustrated in Fig. 1. Firstly, the missing instances can occur because conventional WSOD takes the highest-score proposal and less discriminative objects are ignored. Secondly, the spatially adjacent targets can be grouped into a single proposal that achieves higher classification score, resulting in grouped instances. Lastly, bounding boxes may only cover part of the objects reaching local minima, which can be referred as part domination.

We introduce a novel model designed to address the aforementioned challenges and enhance the detection ca-

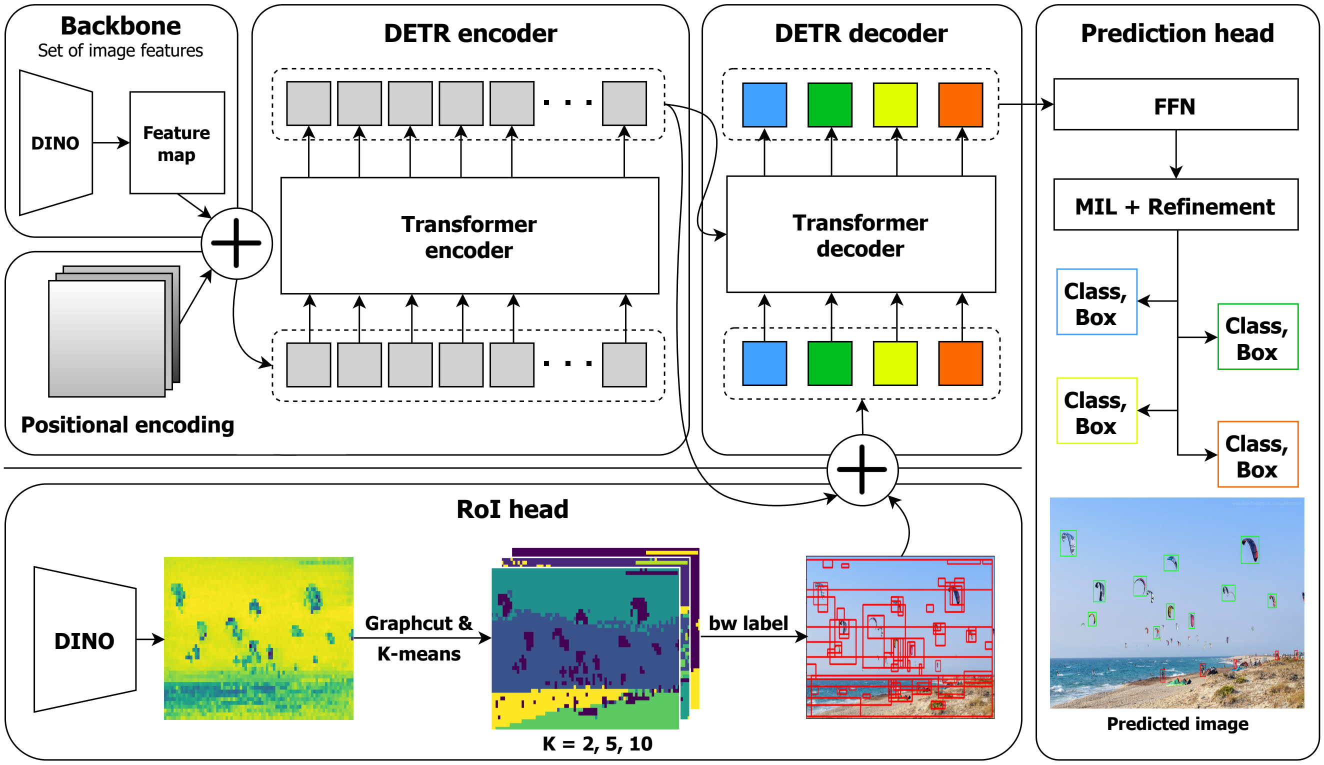

Figure 2. **Our architecture.** To exploit the aforementioned property of self-supervised vision transformer, *i.e. semantic segmentation*, we use the pre-trained vision transformer, DINO [8] as feature extractor. The generation of bounding box proposals is based on the affinity matrix computed from the last attention layer of DINO. Subsequently, Graphcut and K-means clustering techniques are employed. The boxes undergo filtering and mean pooling, leveraging the attention map and similarity score. Once processed, the bounding boxes are then fowarded to the decoder as queries. Within the DETR decoder head, the class score and bounding box predictions are made through FFN, followed by MIL and refinement head.

pacity of scale-variant objects. The proposed architecture consists of several key components: a feature extraction network, an informed proposal generation stage, a WSOD framework based on DETR [7], and proposal refining stage as described in Fig. 2. The feature extraction network plays a pivotal role in capturing instances within the image, regardless of their size. To accomplish this, we leverage DINO [24], which enables the encoder to learn unsupervised representations. Additionally, we incorporate DETR network to further fine-tune it for multi-label classification task. The proposed bounding boxes obtained from the clustered similarity matrix are forwarded to the decoder as queries, implicitly providing the target location information. Considering that the decoder's heads produce predictions for each bounding box, matching the number of queries, we filter the proposed bounding boxes using the similarity and attention score with reference to class token. The resulting outputs from decoder head are subsequently refined at the prediction head, constructed based on the WSDDN-OICR model [42]. The refinement head utilizes the predictions from MIL as pseudo ground truths for the subsequent refining stages.

The paper makes several contributions, which can be summarized as follows:

- Our proposed method utilizes self-supervised feature extractor in the WSOD approach, thereby it detects multi-scale instances, especially small objects.

- We propose an informed proposal generation and IoU thresholding to effectively predict and refine bounding box localization.

2. Related work

Our work has been established based on prior works in various domains: weakly supervised object detection (WSOD), vision transformers in self-supervised manner, and multiple instance approach deploying pseudo groundtruths.

2.1. Weakly-supervised object detection (WSOD).

In recent years, weakly-supervised object detection (WSOD) has shown remarkable performance [6, 22, 35, 41, 42] gaining great attention in the field of computer vision. WSOD is a data-efficient alternative to object detection by being trainied in fully-supervised manner since it only requires image-level labels for training a detector without any other information, *e.g.* bounding-box coordinates. Early works typically formulated the problem of training WSOD as classification task to select the most plausible candidate with high confident given multiple proposals [9, 38, 62]. Follow-up works have been introduced several methods including initialization, representation and regularization techniques to modify this traditional task and showed the improvements [5, 21, 29, 40, 50]. Applying the up-to-date augmentation techniques such as Cutout [56] and CutMix [14] also led to further improvement for booth classification and localization performance. Bilen *et al*. [6] proposed a Weakly Supervised DeepDetection Network (WSDDN), the first MIL(Multiple Instance Learning)-based end-to-end trainable CNN-based model for this task and subsequent works have focussed on generating better pseudo groundtruths [10, 24, 27, 41, 42, 57]. Moreover, to fully deploy the limited information given by weak-supervision condition, Transformer [44] has been introduced to this task and showed great performance [28, 52]. However, since WSOD still suffers well-known chronic problems, *i.e.* focussing on single salient object or discriminative object part and treating clustered instances as single object due to the lack of a formal definition of obejcts, leading to inferior performance to fully-supervised couterparts. In this work, we combined the self-supervised feature extractor to convey instance-wise attention map to WSOD network to mitigate these problems.

2.2. Multiple Instance Approach.

After Multi-Instance Learning (MIL) framework first proposed in [15], which is a task to find instances in the bags where each bag may contain multiple instances, various solutions have been proposed from early improvements [1,50,58] to recent works including employment of pooling mechanisms or self-attention and introduction of generative models [25, 37, 49, 59, 60]. In this task, model is allowed only for the labels of the bags while specific labels of instances are not given. Due to this characteristics, MIL has been widely adopted to solve the WSOD task, by switching the problem to defining an appropriate instance classifier only given labels of bags. MIL-based WSOD works have shown great performances [6,12,26,32,35,55]. To successfully consider multi-instances, along with aforementioned WSSDN, various methods have been proposed including alternative relabelling [12], end-to-end framework combined of CNN and MIL [26, 32], Multiple Instance Self-Training (MIST) [35] and feature bank for additional pseudo label [55].

2.3. Self-supervised vision transformers.

After successful adaptation of transformer architecture [44] to vision domain [17], various self-supervised methods have been attempted with vision transformers to utilize the mass amount of unlabeled data without human-labored dataset. DINO [8] proposed to apply self-distillation loss to vision transformer which leads model to learn useful information for semantic segmentation of images. Inspired by Masked Language Modeling (MIM) method deployed in BERT [13], BEIT [3] first introduced image tokenization and modify the MIM task to Masked Image modeling task, which randomly mask out the image patches and train model to recover those tokens. The follow-up work iBOT [63] incorporated the self-distillation loss from DINO to MIM task, reporting that MIM task enforces model to implicitly learn semantic segmentation, *i.e.* learned representation can partition image patches into groups with clear semantic meaning. Recently, masked autoencoder [23] have shown superior performance in MIM task which trains model via masking out high portion of random image patches and reconstruct the patches in raw image form. Since representations learned in self-supervised manner are not over-focussing on classification label, but only learns general features or semantic segmentations of image, we adopt pre-trained features trained by self-supervision to better detect the diverse objects.

3. Approach

In this section, we introduce our novel approach, as illustrated in Fig. 2. Our approach is based on two novel methods which are using self-supervised technique and informed proposal generation respectively. As with most state-of-the-art WSOD models, the architecture of our model is mainly based on MIL head [6] and refinement head [42]. In the following, we first present these basic methods in Section 3.1 and then introduce our novel techniques in Section 3.2.

3.1. Background

3.1.1 Multiple Instance Learning Head

Multiple Instance Learning (MIL) is a type of supervised learning which refers image as a bag of multiple instances [45] as mentioned in Section 2.2. To address the approach, we adopt the method developed by Bilen and Vedaldiwhich [6].

Initially, the generated proposals and their scores passed through two fully connected layers. The resulting proposal features are then split into two separate streams, each of which uses fc layers to produce two matrices, x_c, $x_d \in \mathbb{R}^{C\times|R|}$. C denotes the number of classes in the image, and

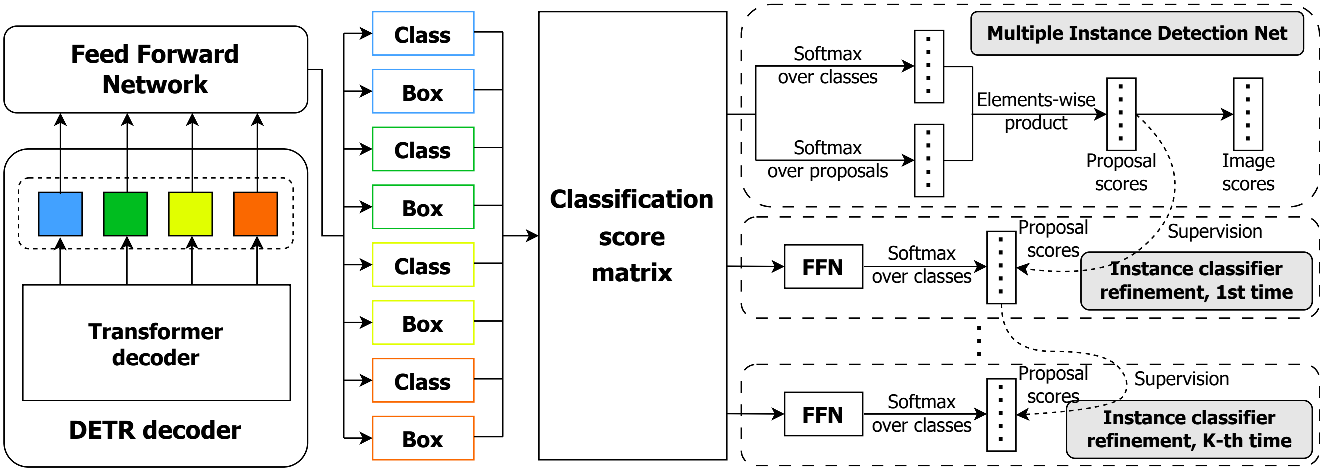

Figure 3. **Bounding box refinement process.** Once the class scores and bounding box locations are predicted from DETR and FFN, the classification scores are concatenated to form a classification score matrix. In the multiple instance detection net, the matrix is computed using softmax over classes and over proposals, followed by element-wise product and summation across classes. The summation is compared with the image label, while the score product serves as the pseudo label for the subsequent instance class refinement stages. By computing the cross-entropy between the proposals and pseudo labels, the predicted bounding box effectively encompasses the whole representation of the target.

$|R|$ represents the number of proposals. Following this, the two matrices are passed through two separate softmax layers along either classes or proposals as outlined in Eq. (1). $[\sigma(x^c)]_{ij}$ corresponds to the probability of proposal j being assigned to class i, while $[\sigma(x^d)]_{ij}$ represents the weighted contribution of proposal j to classify the image as class i. An element-wise product is computed to determine the proposal scores, resulting in $x^R = [\sigma(x^c)] \odot [\sigma(x^d)]$. The summation of the score over proposals ($\Phi_c = \sum_{r=1}^{|R|} x^R_{cr}$) is then compared to the ground truth class label, serving as a basis for the loss function in Eq. (2).

$$[\sigma(x^c)]_{ij} = \frac{e^{x^c{}_{ij}}}{\sum_{k=1}^{C} e^{x^c{}_{kj}}}, [\sigma(x^d)]_{ij} = \frac{e^{x^d{}_{ij}}}{\sum_{k=1}^{|R|} e^{x^d{}_{kj}}} \tag{1}$$

$$L_b = -\sum_{c=1}^{C} \{y_c \log \Phi_c + (1-y_c) \log (1-\Phi_c)\} \tag{2}$$

3.1.2 Refinement Head

The refinement head, in conjunction with MIL architecture, facilitates the iterative adjustment of an instance classifier. The overall architecture is based on the framework proposed by Tang et al. [42].

The r-th candidate proposal at k-th refinement $x_r^(k) \in \mathbb{R}^{(C+1)}$ is generated through an informed proposal generation, which will be further elucidated in Section 3.2.1. Each proposal contains the is represented by a $C+1$-dimensional vector containing class scores for each individual class as well as the background. The label vector for the r-th proposal at the k-th iteration, $y_r^{(k)} \in \mathbb{R}^{(C+1)}$, then stores the highest proposal scores from the previous iteration (k-1) and serves as a supervision for next refinement time k. In each iteration, the assigned class label of an adjacent proposal j, typically representing larger proposals, is determined based on the IoU between the proposal with high score j_c^k and j. If the IoU, I_r, exceeds a predetermined threshold I_t, the assigned class label is set as class c ($y_{cj}^k = 1$); otherwise, it is assignmed as background ($y_{(C+1)j}^k = 1$).

Each refinement process, therefore, aims to gradually identify larger proposal that covers the object. The objective is accomplished by minimizing the the cross-entropy loss in Eq. (3).

$$L_r^k = \frac{1}{|R|} \sum_{r=1}^{|R|} \sum_{c=1}^{C+1} w_r^k y_{cr}^k \log x_{cr}^{Rk} \tag{3}$$

where w_r^k is the loss weight that ensures stability, particularly during initial stages when noisy supervision may be acquired.

And the overall network is trained in Eq. (4) by integrating the respective refinement stages.

$$L = \sum_{k=1}^{K} L_r^k \tag{4}$$

where K is the total number of refinement steps.

3.2. Our Approach

3.2.1 Informed Proposal Generation

Feature Map Generation. Since the model have no access to any information about foreground nor background, there has been several attempts to propose effective proposal generation method [2, 31, 43]. These random bounding box proposals however, make object detection task even harder in WSOD setting. To mitigate this challenge, we suggest to deploy pre-trained feature to generate informed proposals which also could achieve high efficiency. Specifically, we train vision transformers in self-supervised manner and extract the attention map from the last layer. Since the self-supervised vision transformers are not focusing on classification-beneficial features, but simply semantically group the patches [63], images are segmented into objects and backgrounds by attention scores. Attention maps are extracted by class token if class token is involved in the model else summation of attention scores for each token is used for attention map.

Considering class token, $T \in \mathbb{R}^{(N+1)\times d}$ input tokens are processed through the attention layers and from the last attention layer, we are able to extract the attention matrix $A \in \mathbb{R}^{(N+1)\times(N+1)}$, where d and N denote the dimensionality of the feature and number of input patches respectively. Following the standard attention mechanism [44], queries $Q \in \mathbb{R}^{(N+1)\times d}$ and keys $K \in \mathbb{R}^{(N+1)\times d}$ are computed from the input tokens I. Then, dot product between Q and K yields the attention matrix A (Eq. (5)).

$$A = QK^T \tag{5}$$

If the feature extractor contains meaningful class token, *i.e.* attention weights of the class token indicate the segmentation of tokens, A is scaled by $\sqrt{d}$ following the softmax function and the first row of it, discarding the first value of class token, is defined as attention score map S (Eq. (6)).

$$S = (\text{Softmax}(A/\sqrt{d}))_{[1,2:]} \tag{6}$$

On the other hand, if class token is not meaningful, each row of A is summed to single value, *i.e.* attention weights of each token is summed to represent the attention score. Then, as we did in the first case, we discard the class token from it and take the rest of it as attention score map S (Eq. (7))

$$S = (\sum_{i=2} A)_{[2:]} \tag{7}$$

Normalized Cuts and K-means Algorithm. Normalized cut partitions the image by applying spectral clustering algorithm on extracted feature maps from the image. Since we have gained attention score map as aforementioned, we can construct fully connected undirected graph via representing each patch token of the image as a node and relationships among them as edges: (Eq. (8))

$$Ncut(A, B) = \frac{C(A, B)}{C(A, \nu)} + \frac{C(A, B)}{C(B, \nu)} \tag{8}$$

where C stands for function to define the similarity between given two clusters A and B. NCut is known to be NP-hard problem, however, when it comes to bipartition, *i.e.*, clustering into two sub-graphs, it can be reformulated into matrix form and can by solved in the area of eigenvalue system. (Eq. (9))

$$(D - W)x = \lambda Dx \tag{9}$$

W is a fully connected undirected graph which is a $N \times N$ symmetric matrix and from W, we can construct D with $d(i) = \sum_j W_{ij}$, which is a $N \times N$ diagonal matrix. Thus, the eigenvector x corresponding to the second smallest eigenvalue λ becomes the real valued threshold to partition the graph.

Fully Connected Undirected from Self-Supervised ViT. Since DINO [8] produces features containing semantic segmentation of the image, it is natural to adopt it for image partitioning. To generate fully connected undirected graph *i.e.*, symmetric graph, we can think about building similarity matrix of 'query', 'key' and 'value' features. Following the TokenCut [51], as it reports that using 'key' shows the superior performance in the ablation study, we extract 'key' features from the last layer of pre-trained DINO and construct fully connected undirected graph by calculating consine similarity of it (Eq. (10)).

$$W_{ij} = \frac{K_i K_j}{||K_i||_2 ||K_j||_2} \tag{10}$$

Kmeans Algorithm for Finding Multiple Candidate Bounding Boxes. Since we have constructed fully connected undirected graph, we can apply NCut to partition the image. However, the limitation of TokenCut [51] and CutLearn [47] was that they can only produce pre-fixed number of segmentations from the image. This problem originates from NCut itself *i.e.*, it only can bipartite the graph since it classifies the each element of second smallest eigenvector into binary class according to the mean of the eigenvector or just classify them with their sign. To combat this intrinsic problem, we deployed kmeans algorithm to consider various cases of clustering results. Applying kmeans algorithm on eigenvector yields clustered patches and then bwlabel algorithm is used to separate the clustered patches into isolated patche groups *i.e.*, if a cluster consists of multiple patch groups where the patches in the same group are neighbors for each other and the patches in the other groups

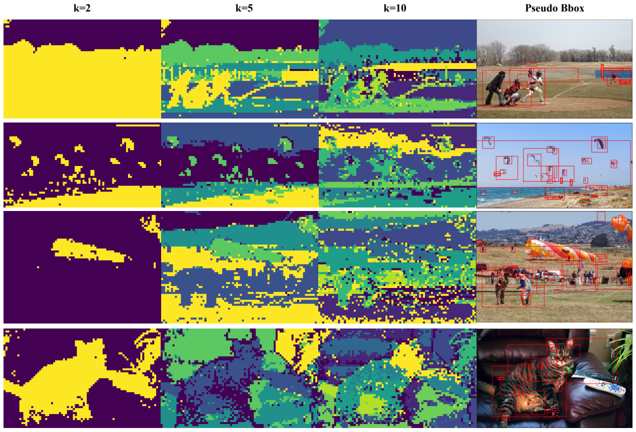

Figure 4. **Examples of segmented images and pseudo bounding boxe proposals.** The K ∈[2,5,10] clusters are obtained from the original image using NCut and kmeans algorithm. Each isolated blob in clusters is processed as single pseudo bounding box candidate. The rightmost column shows the final pseudo bounding box proposals after filtering with IoU.

are cut off. As a default setting, $K \in [2, 5, 10]$ is considered to consider various cases. If K is small, the image will be clustered into relatively large groups containing more patches which is expected to catch large object and on the other hand, if K is big, the image will be clustered into relatively small groups leading to yield more number of small objects.

Attention Score Based IoU Thresholding. IoU thresholding is used during the refinement stage of proposals. In Section 3.1.2, we explained that the refinement head identifies the proposal with the highest score as the anchor, subsequently computing the IoU between the anchor and the rest. Within the process, we have made the specific adjustment: initially filtering the proposals based on a weighted sum of the classification and similarity scores. The attention score is defined as (Eq. (6)) and similarity score is based on the graph constructed from (Eq. (10)) as $S_{sim}(i) = (\sum_j W_{ij})$ where i denotes the order of patch token.

3.2.2 Pseudo Bounding Box as DETR Decoder Query.

DETR Architecture [7]. Our model is built on the framework of DETR [44], a transformer-based architecture designed to process the entire image in a parallel and global manner, effectively capturing long-range dependencies and contextual information. In DETR, the feature map extracted from CNN architecture such as VGG16 or ResNet101, is encoded and utilized as keys and values for decoder. DETR decoder takes positional embeddings as queries, which contains the raw positional information of patches. DETR then directly predicts both labels and bounding boxes in a single unified model, and also eliminate the need for complex region proposal networks (RPN) or anchor-based methods, the common components in conventional object detectors.

Injection of Pseudo Bounding Box Information. Since DETR suggests the Transformer based architecture which does not have bounding box proposal(*e.g.*, selective search [43]) module, but directly predicts through the attention based decoder, we cannot use the bounding box information as usual WSOD models. To fully employ this simple structure, we injected bounding box information into the decoder query. We combined positional embeddings and output tokens from the encoder to consider both of learned features and positional information. Then, patches in each pseudo bounding box are mean-pooled to get query vectors. Since these query vectors are linear combination of learned features composing specific pseudo bounding box, they are expected to be armed with pseudo bounding box information leading to alleviate the problem that we can only use multi-class information while the ground truth bounding boxes are not allowed during training.

3.2.3 Implementation Details

Feature Extractor. As mentioned above, we adopted DINO for our feature extractor and pseudo bounding box annotation generator to exploit its image segmenting features. DINO is pre-trained for 300 epochs on ImageNet-1K for 300 epochs following the orignal recipe [8].

Datasets. We adopted MS COCO 2014 and Pascal VOC 2007 as our datasets. COCO 2014 contains 123K for training, 41K for validation image from 80 classes while VOC 2007 dataset includes 5K for training, 5K for validation images from 20 classes.

Metrics. During training, referring to WSOD frameworks, MIL head and refinement head are used, but we modified the classification part of MIL head to multi-label classification task. For evaluation metric, the primary metrics in object detection task domain *i.e.*, average precision(AP, including AP_S, AP_M, AP_L) and mean average precision (mAP) were used for COCO dataset and PASCAL VOC dataset respectively.

4. Experiments

This section describes how our model was used to conduct numerous experiments on various datasets that are commonly used for object detection. The results were compared with various WSOD architectures to demonstrate the effectiveness of our proposed method.

Table 1. **Comparison of mAP (%) for different state-of-the-art algorithms on VOC 2007 test set.** We report performance the detection accuracy represented by mAPs (%) per class on the VOC 2007 test set. The listed algorithms are either based on CNN or transformer, and utilized VGG16, ResNet50, ResNet101, or the transformer for the feature extractor.

Model	Feature extractor	Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
CNN-based	VGG16	WSDDN [6]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
		OICR [42]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
		C-MIL [54]	**62.5**	58.4	**49.5**	**32.1**	**19.8**	**70.5**	**66.1**	**63.4**	20.0	**60.5**	**52.9**	**53.5**	**57.4**	**68.9**	8.4	**24.6**	**51.8**	**58.7**	**66.7**	**63.5**	**50.5**
		PCL [41]	54.4	**69.0**	39.3	19.2	15.7	62.9	64.4	30.0	**25.1**	52.5	44.4	19.6	39.3	67.7	**17.8**	22.9	46.6	57.5	58.6	63.0	43.5
	ResNet50	WSDDN [6]	50.4	56.7	41.8	24.9	29.9	64.0	55.8	47.8	21.5	50.3	35.0	49.5	49.5	58.1	13.9	24.5	44.7	40.7	65.3	55.8	44.0
		OICR [42]	61.2	50.9	55.0	33.2	36.2	68.6	65.7	**79.2**	17.3	58.1	19.3	69.1	65.7	64.8	15.1	18.9	50.1	55.1	**69.8**	**64.4**	50.9
		C-MIL [54]	**67.5**	45.2	**62.9**	**33.4**	**41.6**	**73.9**	66.7	76.2	**26.4**	54.8	11.6	**71.4**	**71.9**	**72.9**	20.6	**31.9**	42.5	**58.5**	77.1	61.3	53.4
		CASD [24]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**56.8**
		PCL [41]	55.4	**60.7**	50.8	30.1	31.0	69.8	**69.0**	66.6	9.6	**62.0**	**25.0**	56.4	68.2	65.5	**35.7**	28.1	**57.2**	52.9	67.0	54.2	50.8
	VGG16	WSDDN [6]	47.0	58.6	40.4	21.1	28.4	68.4	57.1	46.5	20.1	49.5	35.5	51.8	48.1	55.8	12.2	19.6	45.4	53.8	63.2	58.1	44.1
		OICR [42]	63.2	51.1	51.9	33.7	32.4	67.9	65.0	**78.9**	19.0	59.4	21.9	**70.6**	68.3	64.4	15.2	20.8	49.3	55.3	72.5	**66.6**	51.4
		C-MIL [54]	**66.7**	41.4	**64.7**	**35.5**	**42.2**	**73.7**	67.3	76.3	**23.4**	56.0	12.1	68.7	**74.5**	**75.1**	22.6	**34.1**	43.6	**60.5**	**76.2**	64.2	**53.9**
		PCL [41]	56.5	**65.4**	54.2	27.8	30.2	70.8	**67.5**	74.8	3.2	**60.4**	**56.0**	68.0	70.6	65.4	**35.8**	23.1	**53.1**	53.0	70.7	60.4	53.3
Transformer	VGG16	WSTDN [52]	60.1	**76.8**	59.9	31.7	29.9	73.4	**72.0**	74.9	**29.2**	**64.4**	47.3	41.0	61.6	69.1	33.7	25.0	**57.3**	**61.4**	67.3	58.0	54.7
	DINO [8]	Ours	28.3	45.4	22.5	13.4	10.1	51.2	31.0	32.3	7.3	27.7	20.4	21.0	29.7	36.5	11.8	13.1	29.1	32.6	41.2	33.8	26.9

Table 2. **Comparison of APs for different state-of-the-art algorithms on MS COCO 2014 test set.** We report performance the detection accuracy of AP, AP50, AP75, APs, APm, and APl on the COCO 2014 test set. The listed methods are either based on CNN or transformer, and utilized VGG16, ResNet50, ResNet101, or the transformer for the feature extractor.

Model	Feature extractor	Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
CNN-based	VGG16	MIST [35]	11.4	24.3	9.4	3.6	12.2	17.6
		CASD [24]	**12.8**	26.4	-	-	-	-
		PCL [41]	8.5	19.4	-	-	-	-
		WSOD2 [57]	10.8	22.7	-	-	-	-
		C-MIDN [20]	9.6	21.4	-	-	-	-
	ResNet50	MIST [35]	12.6	26.1	10.8	3.7	13.3	19.9
		CASD [24]	**13.9**	27.8	-	-	-	-
		OD-WSCL [39]	**13.9**	**29.1**	**11.8**	**4.9**	**16.8**	**22.3**
		wetectron [36]	12.6	26.3	-	-	-	-
	ResNet101	MIST [35]	13.0	26.3		-	-	-
		OD-WSCL [39]	**14.4**	**29.0**	**12.4**	**4.8**	**17.3**	**23.8**
		wetectron [36]	13.0	26.3	11.4	3.5	13.7	20.4
Transformer	DINO	MIST [35]	7.4	14.8	6.1	2.4	8.3	13.1

4.1. Pseudo Bounding Box Generation

As described above, our main contribution is heavily associated with quality of pseudo bounding box. As shown in Fig. 4, we consider total 3 variation of K for k-means algorithm to take possibility of various clustering results into account. When $K = 2$, the image is forced to be segmented into 2 clusters. Since image is not guaranteed to only have clear object and homogeneous background, $K = 2$ is used to catch a single large object for the specific cases. On the other hand, $K = 10$ yields cluttered clusters which may lead to separation of discriminative parts from objects. However, due to this cluttered clustering, it is able to catch the small objects(*i.e.*, smaller than 2×2 patches) with enough size of K(*e.g.*, $K = 10$). Then, bwlabel algorithm is applied to discriminate all the isolated patch groups and each of those groups became one single pseudo bounding box. By wrapping up all these pseudo bounding box candidates, we conduct IoU to generate fine-filtered informed pseudo bounding boxes.

4.2. Transformer Based Object Localization

Armed with informed pseudo bounding boxes from DINO, our model reported plausible performance compared to previous WSOD works. One of our main contribution is that we introduces Transformer [44] based object detection model [7] to WSOD domain and reports promising results on COCO2014(Tab. 2) and VOC 2007(Tab. 1). Although our model reports inferior performance to previous WSOD models, it shows that it is possible to adopt transformer-based novel object detection framework(*i.e.*, DETR) for WSOD task. We modified DETR architecture to fit WSOD task by leveraging DINO to generate informed pseudo bounding box information and introducing novel method to inject this generated bounding box information in the middle of the model. As a result, we successfully suggest Transformer based WSOD framework reporting plausible performances on COCO 2014 and VOC 2007, implicating the possibility of further improvement on this framework.

4.3. Small Object Detection in WSOD Manner

We evaluate our model on a subset of COCO dataset, consisting of objects labeled as 'small'. We note that as far as we know, our model is the first one targeting small object detection in WSOD domain. In Tab. 2, previous WSOD models shows relatively weak performance on small object dataset or they are not even reporting experiment results for it. Although our model reports comparatively low performance to previous models, the gap between AP_S and AP is much lower than that of previous works, which means our methods successfully combats the small object ignorance problem and shows scale-invariant object detection performance. This unprecedented performance mainly attributes to the characteristics of our model *i.e.*, fine-grained features from DINO which provides image segmentation information and Ncut with k-means algorithm which generates multi-scale bounding boxes. Then when these methods converge and eventually generates queries which are calculated by mean-pooling the patch vectors in each pseudo bounding box, it intrinsically treats bounding boxes of arbitrary scales to same sized vectors, leading pseudo bounding box information to become scale-invariant query vectors.

Table 3. **Ablation experiments on decoder query and patch size of feature extractor with COCO 2014**. Since we need to consider bounding box information inside the model due to the DETR structure, we proposed to consider it at the decoder step, specifically with the queries. Adding bounding box information to the positional embeddings led to significant performance gain. For patch size ablation study, as smaller patch size leads to more precise image segmentation in pre-training, 8×8 patch size showed superior performance on 16×16 patch size.

Query	COCO 2014 (AP)
Positional encoding (DETR)	3.8
Positional encoding + bbox information (ours)	7.4 (+3.6)
Patch size of feature extractor (DINO)	**COCO 2014 (AP)**
16×16	4.6
8×8	7.4 (+2.8)

4.4. Ablation Studies

We analyze the architecture to demonstrate that our method is effective and explain about our design decisions. We evaluate our model on COCO 2014 and all ablation studies are conducted following our default framework unless otherwise noted.

4.4.1 Pseudo Bounding Box Based Query.

As mentioned in 3.2.2, we exploit the pseudo bounding boxes generated from the DINO for queries. To successfully leverage this information with DETR framework, we need to inject this information in the middle of the model. As a result, we added bounding box information to the query vector of the decoder *i.e.*, positional embedding. In fact, simply using positional embedding as query means we do not use any bounding box information but solely relying on positional embedding to generate bounding box from the decoder. As shown in Tab. 3, our method clearly outperforms the competitor.

4.4.2 Patch Size of Feature Extractor.

Since our proposed method relies on feature extractor(*i.e.*, DINO), we analyze the effect of patch size of DINO. With 16×16 patch size, DINO produces more coarse clustering results which can be led to less noisy pseudo bounding boxes and 16×16 patch size seems enough to catch small objects. However, 8×8 patch size turned out to have superior performance to 16×16 patch size. This result can be attributed to introduction of k-means algorithm to the graph cut, because with diverse size of K, it is able to catch various size of objects and smaller patch size can cover the performance of bigger one.

5. Conclusion

In this section, we provide an overview of the progress made so far and present our future work plan. We note that our main contribution is that we proposed Transformer based WSOD framework and showed the potential of this framework.

Contribution. We suggested novel framework to fully exploit the benefit of Transformer based object detection model(*i.e.*, DETR) and to successfully transplant WSOD structure to it. At the same time, our proposed methods also target small object detection task. As far as we know, both of our trials are the first attempts in WSOD field. DETR has novel structure which does not have bounding box proposal network(*e.g.*, selective search) nor RoI pooling but directly predicts class label and bounding box coordinates from decoder query tokens. Meanwhile, in traditional WSOD frameworks, bounding box proposal is inevitable since the model has to be trained by MIL loss with given features of proposed bounding box and refine this proposed bounding box. To combat this essential disjunction between Transformer based object detection framework and WSOD framework, we proposed informed pseudo bounding box generator and decoder query combined with it. We evaluated our model on COCO 2014 and VOC 2007 to demonstrate our novel idea and the result shows the plausible performance on both of datasets. Most importantly, the gap between AP_S and total AP was remarkably smaller than those of previous WSOD models, indicating our proposed methods work for better small object detection.

Future Work. Considering the plausible performance of our model, it is promising to further improve our architecture and fine-tune targeting SOTA performance. As mentioned above, CIoU [48] and GWD [46] will be applied instead of IoU to produce more precise pseudo bounding boxes however, since applying IoU not only in prediction head but also in bounding box generation module may degrade the performance, ablation study is needed to verify the double usage of IoU. Another part to focus on is how to construct decoder query. While our proposed decoder query works as our hypothesis(*i.e.*, mean vector of pseudo bounding box leads query to attend on whole body of the object), it also leads to group the similar objects nearby. To solve this problem, more studies on decoder query are needed and With improved pseudo bounding box generator our model have lots of potential to be modified. More datasets(*e.g.*, COCO 2017, VOC 2012, etc.) should be tested for evaluation and for fair comparison, previous WSOD models should be re-evaluated with DINO feature extractor.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002. 3

[2] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014. 5

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[4] Aduen Benjumea, Izzeddin Teeti, Fabio Cuzzolin, and Andrew Bradley. Yolo-z: Improving small object detection in yolov5 for autonomous vehicles. *arXiv preprint arXiv:2112.11798*, 2021. 1

[5] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1081–1089, 2015. 3

[6] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2846–2854, 2016. 3, 7

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2, 6, 7

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3, 5, 6, 7

[9] A Chattopadhay, A Sarkar, P Howlader, and Grad-cam++ Balasubramanian. Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. 3

[10] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Slv: Spatial likelihood voting for weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12995–13004, 2020. 3

[11] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *arXiv preprint arXiv:2207.14096*, 2022. 1

[12] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2016. 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[15] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. 3

[16] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7778–7796, 2021. 1

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[18] Shikha Dubey, Farrukh Olimov, Muhammad Aasim Rafique, and Moongu Jeon. Improving small objects detection using transformer. *Journal of Visual Communication and Image Representation*, 89:103620, 2022. 1

[19] Florian Dubost, Hieab Adams, Pinar Yilmaz, Gerda Bortsova, Gijs van Tulder, M Arfan Ikram, Wiro Niessen, Meike W Vernooij, and Marleen de Bruijne. Weakly supervised object detection with 2d and 3d regression neural networks. *Medical Image Analysis*, 65:101767, 2020. 1

[20] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9843, 2019. 7

[21] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2409–2416, 2014. 3

[22] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. Strengthen learning tolerance for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7403–7412, 2021. 3

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[24] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *Advances in neural information processing systems*, 33:16797–16807, 2020. 1, 2, 3, 7

[25] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Inter-*

national conference on machine learning, pages 2127–2136. PMLR, 2018. 3

[26] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 350–365. Springer, 2016. 3

[27] Satoshi Kosugi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object-aware instance labeling for weakly supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6064–6072, 2019. 3

[28] Tyler LaBonte, Yale Song, Xin Wang, Vibhav Vineet, and Neel Joshi. Scaling novel object detection with weakly supervised detection transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 85–96, 2023. 3

[29] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016. 3

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 1

[31] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Convolutional oriented boundaries. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 580–596. Springer, 2016. 5

[32] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015. 3

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[35] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10607, 2020. 1, 3, 7

[36] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[37] Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zielinski. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1721–1730, 2021. 3

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[39] Jinhwan Seo, Wonho Bae, Danica J Sutherland, Junhyug Noh, and Daijin Kim. Object discovery via contrastive learning for weakly supervised object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 312–329. Springer, 2022. 1, 7

[40] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. Weakly-supervised discovery of visual pattern configurations. *Advances in Neural Information Processing Systems*, 27, 2014. 3

[41] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018. 3, 7

[42] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2843–2851, 2017. 1, 2, 3, 4, 7

[43] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104:154–171, 2013. 5, 6

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 5, 6, 7

[45] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019. 3

[46] Jinwang Wang, Chang Xu, Wen Yang, and Lei Yu. A normalized gaussian wasserstein distance for tiny object detection. *arXiv preprint arXiv:2110.13389*, 2021. 8

[47] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023. 5

[48] Xufei Wang and Jeongyoung Song. Iciou: Improved loss based on complete intersection over union for bounding box regression. *IEEE Access*, 9:105686–105695, 2021. 8

[49] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 3

[50] Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai. Relaxed multiple-instance svm with application to object discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1224–1232, 2015. 3

[51] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *arXiv preprint arXiv:2209.00383*, 2022. 5

[52] Zhaofei Wang, Weijia Zhang, and Min-Ling Zhang. Transformer-based multi-instance learning for weakly supervised object detection. *arXiv preprint arXiv:2303.14999*, 2023. 3, 7

[53] Xiwen Yao, Xiaoxu Feng, Junwei Han, Gong Cheng, and Lei Guo. Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):675–685, 2020. 1

[54] Qixiang Ye, Fang Wan, Chang Liu, Qingming Huang, and Xiangyang Ji. Continuation multiple instance learning for weakly and fully supervised object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5452–5466, 2021. 7

[55] Yufei Yin, Jiajun Deng, Wengang Zhou, and Houqiang Li. Instance mining with class feature banks for weakly supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3190–3198, 2021. 3

[56] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[57] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. 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*, pages 8292–8300, 2019. 3, 7

[58] Qi Zhang and Sally Goldman. Em-dd: An improved multiple-instance learning technique. *Advances in neural information processing systems*, 14, 2001. 3

[59] Weijia Zhang. Non-iid multi-instance learning for predicting instance and bag labels using variational auto-encoder. *arXiv preprint arXiv:2105.01276*, 2021. 3

[60] Weijia Zhang, Xuanhui Zhang, Min-Ling Zhang, et al. Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:34940–34953, 2022. 3

[61] Qijie Zhao, Tao Sheng, Yongtao Wang, Feng Ni, and Ling Cai. Cfenet: An accurate and efficient single-shot object detector for autonomous driving. *arXiv preprint arXiv:1806.09790*, 2018. 1

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[63] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3, 5