Read, look and detect: Bounding box annotation from image-caption pairs

Eduardo Hugo Sanchez

Read, look and detect: Bounding box annotation from image-caption pairs

Eduardo Hugo Sanchez
IRT Saint Exupéry
Toulouse, France
eduardo.sanchez@irt-saintexupery.com

Abstract

Various methods have been proposed to detect objects while reducing the cost of data annotation. For instance, weakly supervised object detection (WSOD) methods rely only on image-level annotations during training. Unfortunately, data annotation remains expensive since annotators must provide the categories describing the content of each image and labeling is restricted to a fixed set of categories. In this paper, we propose a method to locate and label objects in an image by using a form of weaker supervision: image-caption pairs. By leveraging recent advances in vision-language (VL) models and self-supervised vision transformers (ViTs), our method is able to perform phrase grounding and object detection in a weakly supervised manner. Our experiments demonstrate the effectiveness of our approach by achieving a 47.51% recall@1 score in phrase grounding on Flickr30k Entities and establishing a new state-of-the-art in object detection by achieving 21.1 mAP_{50} and 10.5 $\text{mAP}_{50:95}$ on MS COCO when exclusively relying on image-caption pairs.

1 Introduction

Locating and classifying objects within an image is a fundamental task in computer vision that enables the development of more complex tasks such as image captioning [47], visual reasoning [21], among others. Nevertheless, the success of object detection models [15, 31] typically relies on human supervision in the form of bounding box annotations. In particular, data annotation is a time-consuming and arduous task that requires annotators to draw bounding boxes around objects and label each bounding box with a category from a fixed set of categories. Furthermore, modifying the number of categories may require annotators to relabel or add new bounding boxes.

Several approaches have been proposed to reduce the cost of data annotation in object detection by using image-level labels [2, 5, 40, 56, 10, 53, 32, 13], a dataset containing both labeled and unlabeled data [37, 24, 25] or sparsely-annotated data [46, 54, 44, 18]. For instance, WSOD methods only use image-level annotations along with the multiple instance learning (MIL) [28] approach. However, the annotation effort is still significant and similar to that required for supervised classification.

In this paper, we take a step forward by learning to locate and label objects within an image from image-caption pairs. Not only captions provide a more natural description of the image content than image-level labels but also constitute a form of weaker supervision since image-caption pairs are easier to collect in vast amounts (e.g. from the Web [34]). Our approach combines recent advances in vision-language (VL) models [19] and self-supervised vision transformers (ViTs) [3].

VL models leverage large-scale image-caption datasets and have strong performance on zero-shot image classification, image-text retrieval, and visual reasoning tasks. These models align images with their corresponding captions via contrastive learning. Notably, models that include a cross-modality

Preprint. Under review.

Figure 1: Our approach leverages language from captions via ALBEF to annotate multiple objects per image (e.g. we detect a dog and a frisbee while LOST generates a single categoryless bounding box).

encoder seem to implicitly learn a more fine-grained word-region alignment without using additional supervision [19]. We propose to use the location ability of VL models to automatically annotate objects of interest mentioned in captions. Moreover, VL models do not require retraining when the number of categories to annotate changes as they are already included in the VL model's vocabulary.

Despite the strong ability of VL models to locate objects, they aim at the most distinctive part of the object rather than the whole object. For example, ALBEF [19] and VilBERT [26] perform phrase grounding by ranking the object proposals provided by the supervised detector MattNet [51]. On the other hand, recent work has shown that representations from self-supervised ViTs contain explicit information about the scene layout of images and produce heatmaps that highlight salient objects [3]. LOST [36] and TokenCut [45] show the effectiveness of self-supervised ViT representations to perform unsupervised object discovery and detection without any labels.

We make the following contributions in this work. First, we propose a novel method to locate and label objects in images by combining the ability of VL models to point at objects and the ability of self-supervised ViTs to extract whole objects in Section 3. By building upon ALBEF [19] and LOST [36], our method is able to locate multiple objects and generate accurate bounding boxes without human supervision. Figure 1 illustrates the improved ability of our model over LOST. Second, we use our approach to perform phrase grounding and object detection in a weakly supervised fashion. In Section 4, we demonstrate that our method achieves competitive performance in phrase grounding on Flickr30k Entities [29] and establish a new state-of-the-art in object detection on MS COCO [23] when exclusively relying on image-caption pairs as unique source of supervision. Additionally, we perform ablation experiments to investigate the key components of our approach, transfer learning experiments on PASCAL VOC2007 [7] and pseudo-labeling experiments to improve the performance in WSOD. In Section 5, we discuss the limitations, future work and conclusions of our work.

2 Related work

Weakly supervised object detection: To reduce the cost of data annotation, several methods propose to train object detectors using only image-level annotations without the need for bounding box annotations. WSDDN [2] introduces the first end-to-end WSOD framework that adopts MIL [28]. Since then, several improvements have been proposed: PCL [40] performs clustering to improve object proposals and W2F [56] leverages pseudo-label mining from a WSOD model to train a supervised object detector. C-MIDN [10] introduces a method for coupling proposals to prevent the detector from capturing the most discriminative object part rather than the whole object. WSOD^2 [53] performs pseudo-label mining and incorporates a bounding box regressor to fine-tune the location of

each proposal. Likely, MIST [32] performs pseudo-label mining where highly overlapping proposals are assigned to the same label. CASD [13] combines self-distillation with multiple proposal attention maps generated via data augmentation. Closely related to our work, Cap2Det [48] learns from image-caption pairs by extracting image-level annotations from captions using a supervised text classifier. These predicted image-level annotations are subsequently used to train a WSOD model based on MIL. Additionally, Cap2Det [48] refines the WSOD model by retraining on instance-level pseudo-labels multiple times. Most of the existing WSOD methods rely on object proposal algorithms (e.g. Selective Search [41] or Edge Boxes[58]). By exclusively leveraging self-supervision on image-caption pairs, our approach outperforms the state-of-the-art model Cap2Det [48] without the need for a supervised text classifier. Additionally, our approach outperforms relevant WSOD baselines [40, 10] that use a form of stronger supervision (image-level annotations) and object proposal algorithms.

Learning from unlabeled or partially labeled data: Some approaches alleviate the lack of bounding box annotations by leveraging a small labeled dataset and a large unlabeled dataset via semi-supervised learning [37, 24, 25] and active learning [43, 42]. Li et al. [18] propose to train an object detector using only a single instance annotation per category per image. Other methods combine image-level and instance-level pseudo-annotations during training [46, 32]. Sohn et al. [37] propose a two-stage training in which an object detector is trained on available labeled data. This model is subsequently used to select high-confidence bounding boxes on unlabeled data as pseudo-labels. Wang et al. [44] address the missing annotation problem by introducing a siamese network where each branch is used to generate pseudo-labels for each other. Likely, recent work [24, 25] leverage the teacher-student framework in object detection. In this work, we also explore the use of pseudo-labels to improve WSOD performance.

VL models: Learning joint VL representations from image-caption pairs in a self-supervised fashion has proven to be effective to perform multiple downstream tasks [26, 27] such as visual question answering, image retrieval, image captioning, zero-shot classification, etc. VL models [39, 20, 38, 16, 8, 22, 4, 14, 30, 19] are generally trained on a combination of loss functions: masked language modelling (MLM), where a masked word token is predicted; masked image modelling (MIM), where a masked image region feature or object category is predicted, image-text contrastive learning (ITC), where positive/negative image-caption pairs are assigned to high/low similarity scores, respectively; and image-text matching (ITM), that predicts whether an image and a caption match. Many strategies have been proposed to achieve improved VL representations. VisualBERT [20] uses a supervised object detector to extract visual embeddings. VILLA [8] performs adversarial training in the representations space. OSCAR [22] uses object tags to ease VL alignment. UNITER [4] encourages alignment between words and image regions extracted by an object detector. More recently, CLIP [30] leverages a massive amount of image-caption pairs and achieves impressive performance at zero-shot classification. However, CLIP underperforms at other VL tasks as the interaction between vision and language is very shallow (i.e. a simple dot product). Li et al. [19] propose a new model called ALBEF, which builds upon previous models [39, 26, 30, 16, 14] and is composed of a vision encoder, a language encoder, and a cross-modality encoder for deeper VL interaction. By leveraging a large image-caption dataset [34], ALBEF outperforms previous models at many VL tasks without the need for a supervised object detector to extract region-based image representations. Our approach leverages a pre-trained ALBEF model to locate the image region that corresponds to a word or a textual description.

Open vocabulary detection: Classifying an object or image has been traditionally limited to a small set of fixed categories. Zhang et al. [55] leverages the vocabulary from image-caption datasets to perform image classification across more than 30k classes. The recent success of VL models [30, 19] has motivated other methods to leverage image-caption pairs to perform object detection on a larger number of categories. Zareian et al. [52] use bounding box annotations from base classes to perform correctly in target classes mentioned in captions. Gao et al. [9] use a supervised object detector trained on MS COCO [23] to generate pseudo-bounding box annotations for categories mentioned in captions. Similar approaches [57, 35] have been proposed by extending CLIP [30]. We also leverage VL models to annotate objects using an arbitrary number of categories in a self-supervised manner without relying on bounding box annotations like previous methods.

Object discovery: Recently, several studies explore methods for object localization that rely solely on visual cues. LOST [36] extracts image representations via a self-supervised ViT [3] which are

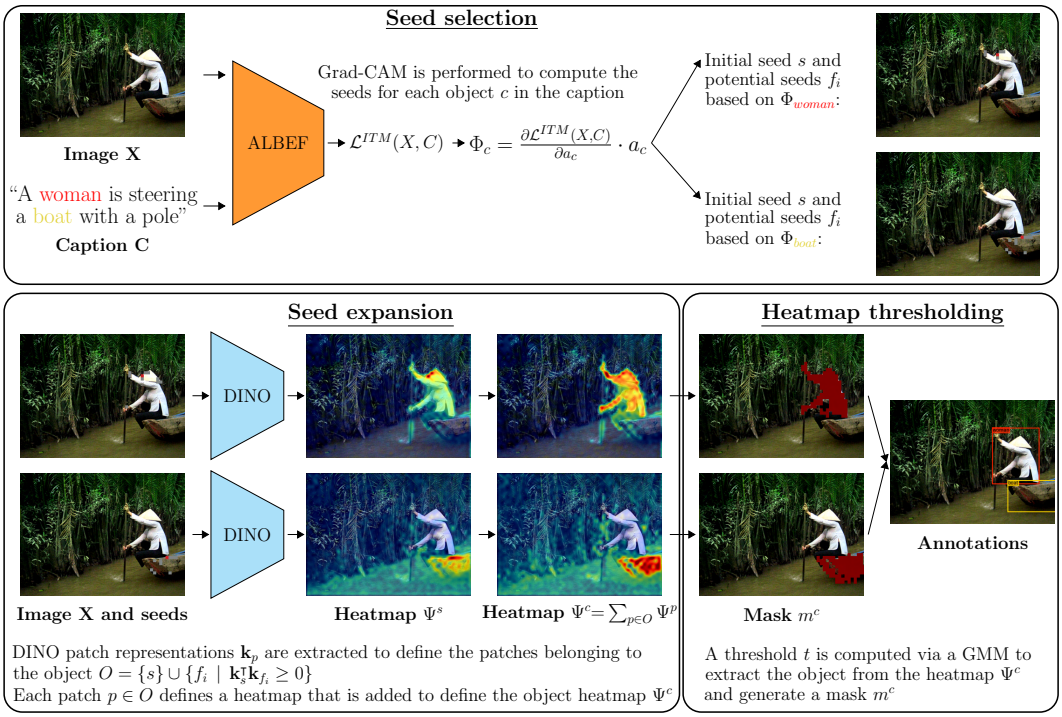

Figure 2: Bounding box generation using the caption "A woman is steering a boat with a pole". First, we select the *initial* and *potential seeds* (red and gray patches, respectively) via a VL model for each category identified in the caption. Second, we perform *seed expansion* by measuring similarity between patches via a ViT. Finally, each heatmap is thresholded and a bounding box is drawn on top.

subsequently used to identify the image patches corresponding to an object based on their correlation. Wang et al. [45] also leverage DINO representations which are used to build a graph. A normalized graph-cut is used to split the foreground object from the background. Both methods can only locate a single object per image without providing its category. Our approach builds upon LOST by integrating the language modality, enabling it to locate and label multiple objects per image.

3 Method

To annotate objects from image-caption pairs, our approach consists of two main stages. First, we leverage the cross-modality encoder from a pre-trained VL model to automatically select the image patches (or *seeds*) that may belong to a given object (defined by a word token or a set of word tokens). The *seed selection* process is described in Section 3.1. Second, we use a self-supervised ViT to compute the similarity between image patches. Intra-image similarity is used to filter out image patches selected in the first stage and generate a heatmap corresponding to the object. This process is known as *seed expansion*. Then, a heatmap threshold is computed via a Gaussian mixture model (GMM) to separate the object patches from the background ones. Finally, a bounding box enclosing the object patches is generated. Section 3.2 describes the process to generate a heatmap and extract an object from it. Figure 2 shows an overview of our approach.

3.1 Pointing at objects with VL models

Our proposed method is motivated by the observation that VL models implicitly learn to align words in the captions with patches in the images even though these models are only trained to align images with their corresponding captions [19]. Furthermore, we can annotate a large amount of objects since the number of objects categories is as large as the vocabulary used in the captions during training of VL models. We leverage the ability of VL models to point at objects and the fact that most of the salient objects in an image are mentioned in its respective caption [22].

In this section, we explain how the fine-grained alignment between words and patches is computed in VL models implementing a cross-modality encoder (e.g. ALBEF [19]) and how we leverage it to point at objects in an image. Let $X = \{p_1, p_2, \ldots, p_{N_P}\}$ be an image composed of N_P patches and $C = \{w_1, w_2, \ldots, w_{N_T}\}$ be its corresponding caption composed of N_T word tokens. An image encoder and text encoder are used to extract image and text representations which are both fed into the cross-modality encoder. In the l^{vl}-th cross-attention layer of this encoder, we compute the value and key representations for each image patch, i.e. $V = \{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{N_P}\}$ and $K = \{\mathbf{k}_0, \mathbf{k}_1, \ldots, \mathbf{k}_{N_P}\}$, respectively, where $\mathbf{v}_0$ and $\mathbf{k}_0$ are the representations of the classification token [CLS].

Given a word token of interest w_c (e.g. 'person', 'dog', etc.), we compute its query representation $\mathbf{q}_c$. The relation between the word token w_c and the image patches $\{p_i\}_{i=1}^{N_P}$ is given by the hidden representation $\mathbf{h}_c$ as shown in Equation 1 where d is the dimension of the query representations.

$$\mathbf{h}_c = \sum_{i=0}^{N_P} a_{c,i} \cdot \mathbf{v}_i \quad \text{where } a_{c,i} = \frac{\exp(\mathbf{q}_c^\intercal \mathbf{k}_i / \sqrt{d})}{\sum_{j=0}^{N_P} \exp(\mathbf{q}_c^\intercal \mathbf{k}_j / \sqrt{d})} \tag{1}$$

As observed, the hidden representation of w_c is a linear combination of the value representations corresponding to the image patches. Furthermore, these representations are weighted according to the attention scores $a_{c,i}$ that implicitly provide the similarity between w_c and p_i via the product $\mathbf{q}_c^\intercal \mathbf{k}_i$. Through the use of the cross-modality encoder, one can identify the image regions that are most closely aligned with a particular word token. We use Grad-CAM [33] to rank the image patches in order of importance. Equation 2 displays the importance score of the image patch p_i with respect to the word token w_c where $\mathcal{L}^{ITM}(X, C)$ is the binary cross-entropy loss that measures whether the image X and the caption C match or not. When ranking image patches, we do not take into account the attention score corresponding to the classification token [CLS], $a_{c,0}$.

$$\Phi_{c,i} = \frac{\partial \mathcal{L}^{ITM}(X, C)}{\partial a_{c,i}} \cdot a_{c,i} \tag{2}$$

Unfortunately, Grad-CAM scores are insufficient to generate an accurate bounding box by themselves (see Section 4). For example, Gao et al. [9] use a supervised Mask R-CNN [12] to generate bounding boxes that cover the activated image patches by the word token w_c for object detection. Similarly, Li et al. [19] rank MattNet [51] proposals based on Grad-CAM maps for phrase grounding.

However, we observe that while Grad-CAM scores do not highlight the image patches corresponding to the whole object, they are useful to point at the most discriminative parts of it. Therefore, we propose to use a set $D = \{f_i\}_{i=1}^{M}$ of M image patches with the highest score $\Phi_{c,i}$ for a given word token of interest w_c to point at the object. The image patches in D are referred to as *potential seeds* and this process is referred to as *seed selection*. Pointing is a natural way for humans to refer to an object [1] and constitutes the first stage of our proposed approach.

3.2 Extracting objects with self-supervised ViTs

We make use of the self-supervised ViT capability [3] to measure the similarity between image patches. Using the location information provided in the previous stage, our approach takes advantage of the fact that object patches correlates positively with each other but negatively with background patches. This idea is successfully applied in LOST [36] to perform object discovery. Our work is inspired by LOST and extends its capabilities by incorporating the language modality.

Assuming that the object area is smaller than the background area, LOST uses the patch with the smallest number of positive correlations with other patches in order to point at an object. However, this assumption may not always hold in practice (e.g. an object covering more area than the background, multiple objects, etc.). Compared to LOST, our method is able to generate multiple bounding boxes per image (as many objects as mentioned in the caption). Furthermore, our method can annotate each object with a label while LOST can only retrieve a single object without specifying its category. Figure 1 displays the differences between our approach and LOST.

In this work, we average the first N patch locations with the highest value of $\Phi_{c,i}$ in D to compute the *initial seed* s for a given w_c. Following LOST, we extract the key representations of the initial and potential seeds, i.e. $\mathbf{k}_s$ and $\{\mathbf{k}_{f_i}\}_{i=1}^{M}$, respectively, from the l^{vit}-th self-attention layer of a self-supervised ViT [3]. Then, the similarity between the initial seed and potential seeds is computed via the dot product of their respective representations to determine the image patches belonging to the

object. We assume that potential seeds that are positively correlated to the initial seed belong to the object while potential seeds that are negatively correlated to the initial seed belong to the background.

Thus, patches belonging to the object are defined by the set $O = \{s\} \cup \{f_i \mid f_i \in D \text{ and } \mathbf{k}_s^\intercal \mathbf{k}_{f_i} \geq 0\}$. Each patch $p \in O$ generates a heatmap $\Psi^p \in \mathbb{R}^{N_P}$, where the i-th dimension Ψ_i^p is computed via the dot product between its key representation $\mathbf{k}_p$ and the key representation of the patch p_i (also extracted by the ViT), i.e. $\mathbf{k}_{p_i}$ $\forall i \in \{1, \ldots, N_P\}$ as shown in Equation 3.

$$\Psi_i^p = \mathbf{k}_p^\intercal \mathbf{k}_{p_i} \tag{3}$$

Finally, the heatmap of the object w_c is defined by the sum of the heatmaps corresponding to the patches in O as shown in Equation 4. This process is referred to as *seed expansion*.

$$\Psi^c = \sum_{p \in O} \Psi^p \tag{4}$$

To extract the object from the heatmap Ψ^c, we define a threshold t. While LOST sets t=0, we assume that patches belonging to the object and background are defined by two normal distributions p_o=$\mathcal{N}(\mu_o,\, \sigma_o^2)$ and p_b=$\mathcal{N}(\mu_b,\, \sigma_b^2)$, respectively. The parameters $\mu_o,\ \sigma_o,\ \mu_b,\ \sigma_b \in \mathbb{R}$ are estimated via a GMM per heatmap with k=2 components. Then, the threshold is calculated by solving $p_o(t)$=$p_b(t)$ such that $\mu_b < t < \mu_o$. For small objects, p_o is barely noticeable and hard to estimate via GMM since only the background component is recognizable. We assume only one component is distinguishable when the overlapping between the estimated distributions p_o and p_b is significant (i.e. $\mu_b + 1.5\sigma_b < \mu_o - 1.5\sigma_o$). In such a case, we use the threshold $t = \mu + \gamma\sigma$ where γ is a constant and μ and σ are the mean and the standard deviation of Ψ^c, respectively. Supplementary material provides bounding box examples using multiple t values. To generate a bounding box, a mask m^c is obtained by thresholding the heatmap Ψ^c as shown Equation 5 where Ψ_i^c is the i-th dimension of the heatmap Ψ^c. Later, a bounding box is drawn by enclosing the segment that includes the initial seed s.

$$m_i^c = \mathbb{1}_{\Psi_i^c \geq t} \tag{5}$$

4 Experiments and results

4.1 Setup details

Tasks and datasets: We perform weakly supervised phrase grounding and object detection to demonstrate the effectiveness of our method to annotate objects. In Section 4.2, we present our experimental results for phrase grounding on Flickr30k Entities [29], an extension of Flickr30k [49] which consists of $\approx$ 32k images collected from Flickr each of which is described with 5 captions. Image-caption samples are split into $\approx$ 30k training, 1k validation, and 1k test samples. Flickr30k Entities includes manually-annotated bounding boxes that are linked with entities mentioned in captions. Results are reported in terms of recall@1 on the test set. In Section 4.3, we perform WSOD on MS COCO [23] which contains 113k training and 5k validation images. Each image is described with 5 captions. Additionally, the dataset provides bounding box annotations covering 80 object categories such as person, bicycle, car, plane, etc. We also conduct transfer learning experiments using samples from MS COCO to train an object detector that predicts PASCAL VOC2007 [7] categories since this dataset does not provide captions. PASCAL VOC2007 is an object detection dataset that contains 2501 training, 2510 validation, and 4952 test images. Objects are labeled into 20 classes (e.g. person, bird, cat, cow, dog, etc.). Results are reported in terms of mean average precision at IoU=0.5, i.e. mAP_{50}, and average mAP over multiple IoU values ranging from 0.5 to 0.95 with a step of 0.05, i.e. $\text{mAP}_{50:95}$. Results are reported on the MS COCO validation set and the PASCAL VOC2007 test set. In all cases, bounding box annotations are only used during evaluation.

Model architecture: To point at objects, we use ALBEF pre-trained on 14M image-caption pairs [19] and fine-tuned on 20k image-caption pairs [50]. It is worth mentioning that any VL model that includes a cross-modality encoder can be used. To perform seed expansion, we use the self-supervised ViT from DINO (i.e. ViT-S/16 [3]). For comparative purposes, we also use the image encoder from ALBEF (i.e. ViT-B/16 [6]) to compute the similarity between image patches. In WSOD, our approach generates bounding box annotations to train a YOLOv5 object detector [15] in a supervised manner.

Hyperparameters: We set the VL cross-attention layer to l^{vl}=8 and the ViT self-attention layer to l^{vit}=11. To compute the initial seed, we average the first N=3 patch locations from D and set the number of potential seeds to M=10. To compute the threshold, we use $\gamma = 1.75$. Our experiments are executed on a NVIDIA GeForce RTX 3090.

Table 1: Weakly supervised phrase grounding performance on Flickr30k Entities.

Method	Training data	Supervised object proposal generator?	Recall@1
ALBEF C-A maps	14M image-caption pairs [19]	No	36.86
ALBEF ViT maps	14M image-caption pairs [19]	No	43.97
DINO ViT maps	14M image-caption pairs [19] + ImageNet images [3]	No	**47.51**
InfoGround [11]	Flickr30k Entities [29]	Yes, Faster R-CNN [31]	47.88
InfoGround [11]	MS COCO [23]	Yes, Faster R-CNN [31]	51.67

4.2 Weakly supervised phrase grounding

We conduct experiments on Flickr30k Entities to evaluate the ability of our approach to associate phrases describing objects to image regions. While a single word can define the category of an object, a phrase provides additional attributes (e.g. color, size, position, etc.). Our method processes phrases by simply adding up the heatmaps of each word w_{c_i} in the phrase P, i.e. $\Psi^{\text{phrase}} = \sum_{c_i \in P} \Psi^{c_i}$.

In Table 1, we report our results in terms of recall@1 that represents the ratio of the number of phrases whose ground truth bounding boxes have significant overlap with the generated bounding boxes by our model (i.e. IoU ≥ 0.5) to the total number of phrases.

Our baseline model (referred to as ALBEF C-A maps) uses the cross-modality encoder to produce heatmaps Φ^{phrase} , which are then thresholded to generate bounding boxes. As shown in Section 3.1, our approach builds upon Φ^{phrase} via a self-supervised ViT to generate the expanded heatmaps Ψ^{phrase}. We evaluate two variants of our approach by using the ViT from ALBEF and DINO to generate the object heatmaps (referred to as ALBEF ViT maps and DINO ViT maps, respectively).

As observed, the variants ALBEF ViT maps and DINO ViT maps achieve higher performance compared to the baseline (improvements of 7.11% and 10.65%, respectively). As hypothesized, the baseline model exhibits limitations in accurately capturing the spatial extent of objects despite its ability to point at them in the image as shown in Figure 3. Moreover, DINO ViT maps outperform ALBEF ViT maps by a margin of 3.54%. This difference suggests that DINO's loss function is more effective to capture the underlying relationships between image patches.

For the sake of comparison, we also report the performance of the state-of-the art model for weakly supervised phrase grounding, i.e. InfoGround [11]. Our approach achieves a competitive score of 47.51% comparable to InfoGround performance (47.88% and 51.67% when trained on Flickr30k Entities and MS COCO, respectively). Nevertheless, InfoGround uses a Faster R-CNN [31] pre-trained on Visual Genome [17] to generate object proposals and extract object features. Thus, our approach offers an efficient solution for phrase grounding without the need for an object detector. Our approach represents a promising alternative to InfoGround, particularly in scenarios where the object detector does not include some categories or where obtaining bounding box annotations is difficult.

4.3 Weakly supervised object detection

We investigate the ability of our approach to perform WSOD. Our methodology involves defining a set of object categories and searching through captions to identify if any of these categories are mentioned. If a category is found, our approach generates a corresponding bounding box as described in Section 3. Then, we train an object detector (i.e. Yolov5 [15]) from scratch in a supervised manner using the generated bounding box annotations. We evaluate our approach on MS COCO [23] and PASCAL VOC 2012 [7]. While our method is capable of labeling a large number of object categories, we use these datasets as they provide bounding box annotations for evaluation purposes.

Comparison with WSOD methods: We compare our approach with state-of-the-art WSOD methods to demonstrate its effectiveness in Table 2. Our approach achieves 21.1 mAP_{50} and 10.5 $\text{mAP}_{50:95}$ on MS COCO outperforming the variants of Cap2Det [48] that learn from image-caption pairs: $\text{Cap2Det}^{\text{EM}}$ that generates image-level annotations from captions via lexical matching and $\text{Cap2Det}^{\text{CLSF}}$ that employs a supervised text classifier to process captions and extract image-level

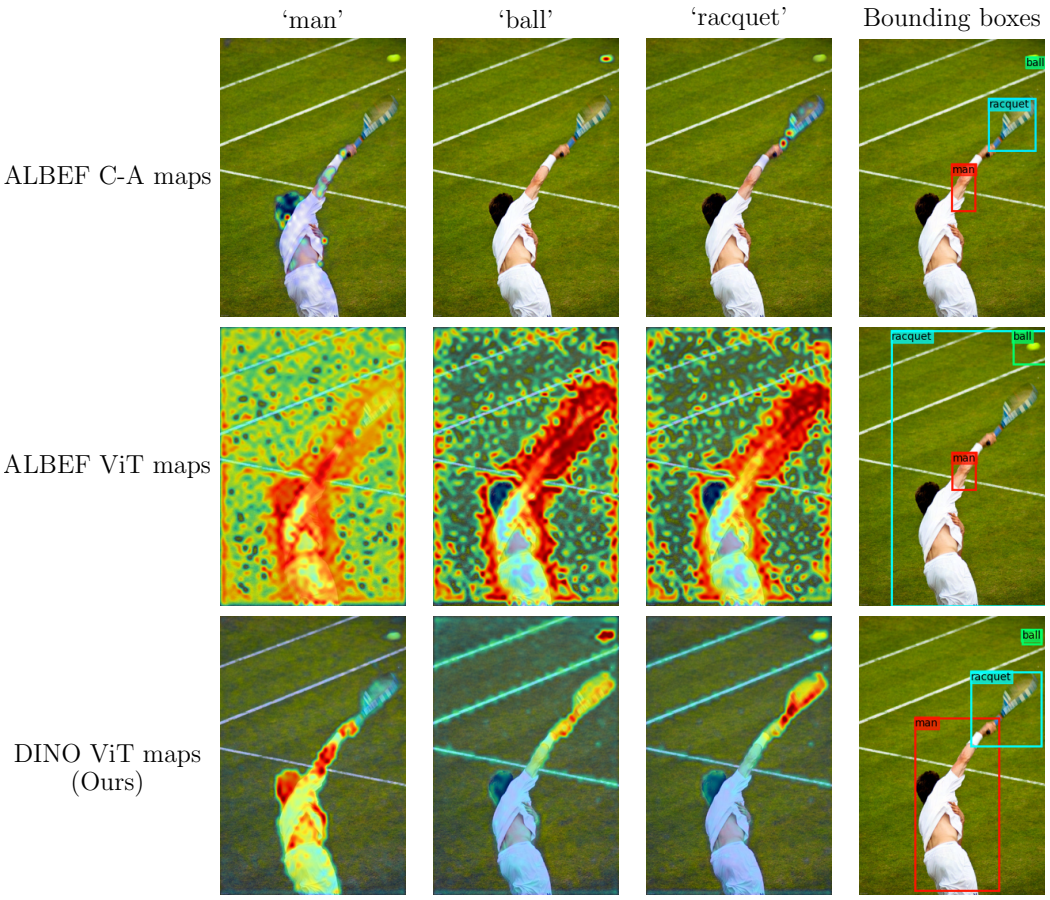

Figure 3: Heatmaps and generated bounding boxes corresponding to 'man', 'ball' and 'racquet'. ALBEF C-A maps point successfully at objects while struggle to get the object extent. ALBEF ViT maps tend to be noisier than DINO ViT maps which generate high-quality bounding boxes.

annotations. Our approach demonstrates better performance without the need for an object proposal algorithm, a supervised text classifier or using refinement. Compared to methods that learn from image-level annotations [10, 53, 32, 13], our approach demonstrates competitive performance and outperforms relevant baselines such as PCL [40] and C-MIDN [10] (8.5 $\text{mAP}_{50:95}$ and 9.6 $\text{mAP}_{50:95}$, respectively) by achieving 10.5 $\text{mAP}_{50:95}$. It is worth noting that these WSOD methods rely on pseudo-labeling techniques and image-level annotations that constitute a form of stronger supervision. For the sake of comparison, we also report the results of Yolov5 trained in a fully-supervised manner.

Transfer learning and pseudo-labeling (P-L): Due to the lack of captions in PASCAL VOC2007, our approach generates annotations by searching PASCAL VOC2007 object categories from MS COCO image-caption pairs. Results in terms of mAP_{50} per category are reported in Table 3 where best results are highlighted in bold. Our approach achieves 40.9 mAP_{50} outperforming $\text{Cap2Det}^{\text{EM}}$ (39.9 mAP_{50}) while being behind $\text{Cap2Det}^{\text{CLSF}}$ (43.1 mAP_{50}). To further improve performance, we propose a simple pseudo-labeling (P-L) technique. First, we use the trained object detector to generate predictions on the training images of PASCAL VOC2007. Pseudo-labels are selected by setting the confidence and IoU thresholds to 0.2 and 0.5, respectively in the NMS algorithm. Then, we fine-tune our trained object detector on these pseudo-labels. We report an improvement of 1.6 mAP_{50} and 1.1 $\text{mAP}_{50:95}$. Despite the global mAP_{50} being inferior to that of $\text{Cap2Det}^{\text{CLSF}}$, it is worth noting that our approach implementing P-L outperforms $\text{Cap2Det}^{\text{CLSF}}$ in many categories.

Ablation experiments: We also perform ablation experiments to identify the key components of our approach in WSOD. To annotate objects, we employ the variants of our approach presented

Table 2: Comparison with WSOD models on MS COCO.

Model	Supervision source	mAP_{50}	$mAP_{50:95}$
$Cap2Det^{EM}$ [48]	image-captions pairs	19.7	8.9
$Cap2Det^{CLSF}$ [48]	image-captions pairs	20.2	9.1
Ours	image-captions pairs	**21.1**	**10.5**
PCL [40]	image-level annotations	19.4	8.5
C-MIDN [10]	image-level annotations	21.4	9.6
$WSOD^2$ [53]	image-level annotations	22.7	10.8
MIST [32]	image-level annotations	25.8	12.4
CASD [13]	image-level annotations	26.4	12.8
Fully supervised [15]	bounding box annotations	66.2	46.7

in Section 4.2, i.e. ALBEF C-A maps, ALBEF ViT maps and DINO ViT maps. Tables 4 and 5 display the results of our experiments on MS COCO and PASCAL VOC2007, respectively. As observed, ALBEF C-A maps perform poorly at object detection achieving the lowest scores mAP_{50} and $mAP_{50:95}$. While ALBEF C-A maps are able to accurately point at objects, they fail to correctly detect their extent. On the other hand, self-supervised ViTs (ALBEF ViT maps and DINO ViT maps) are effective to capture the extent of objects through the seed expansion. In MS COCO, DINO ViT maps outperform ALBEF ViT maps as expected since DINO ViT maps are less noisy and generates visually more accurate bounding boxes as shown in Figure 3. Surprisingly, ALBEF ViT maps achieve slightly better results than DINO ViT maps in PASCAL VOC2007.

Table 3: Comparison with WSOD models on PASCAL VOC2007.

Model	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP_{50}
$Cap2Det^{EM}$ [48]	63.0	50.3	50.7	25.9	**14.1**	64.5	**50.8**	33.4	**17.2**	49.0	48.2	46.7	44.2	59.2	10.4	14.3	49.8	37.7	21.5	47.6	39.9
$Cap2Det^{CLSF}$ [48]	**63.8**	42.6	50.4	29.9	12.1	61.2	46.1	41.6	16.6	**61.2**	**48.3**	**55.1**	**51.5**	59.7	16.9	15.2	50.5	**53.2**	38.2	**48.2**	**43.1**
Ours	58.8	64.6	52.3	28.9	10.0	57.2	42.2	**50.7**	12.8	54.3	32.4	38.8	37.4	**61.9**	**24.2**	17.6	47.3	39.0	52.3	34.4	40.9
Ours + P-L	56.1	**68.5**	**55.6**	**31.1**	12.3	**64.8**	48.6	48.8	15.5	57.8	22.9	34.8	42.3	59.1	23.2	**19.1**	**51.8**	42.8	**54.8**	41.0	42.5
Supervised [15]	70.2	74.3	42.8	40.4	40.8	73.6	83.3	62.0	37.7	61.3	58.3	56.1	77.5	71.2	78.0	35.5	50.5	55.0	75.1	60.2	60.2

Table 4: Ablation experiments on MS COCO.

Method	mAP_{50}	$mAP_{50:95}$
ALBEF C-A maps	9.4	3.7
ALBEF ViT maps	18.4	9.0
DINO ViT maps	**21.1**	**10.5**

Table 5: Ablation experiments on VOC2007.

Method	mAP_{50}	$mAP_{50:95}$
ALBEF C-A maps	9.2	3.3
ALBEF ViT maps	**42.9**	**20.8**
DINO ViT maps	40.9	18.0

5 Conclusion

In this paper, we present a two-stage method to locate and label objects by leveraging image-caption pairs without additional supervision. We demonstrate the effectiveness of our approach by performing two tasks in a weakly supervised setting: phrase grounding and object detection. We have performed extensive experiments on Flickr30k Entities, MS COCO and PASCAL VOC2007 achieving state-of-the-art results without the need for supervised object proposal algorithms or text classifiers to process captions. Despite the remarkable performance of our approach, we acknowledge some limitations. Our approach produces a single bounding box per object mentioned in the caption. An interesting direction for further investigation is to extend our method to produce multiple bounding boxes for words representing more than one object instance in the image (e.g. "people", "group of animals", etc.). This is particularly challenging, especially when object instances are overlapping in the image. Also, our approach does not generate bounding boxes for objects present in the image but not mentioned in the caption (or due to spelling mistakes). We believe that an important direction for future work is to extend our approach to explicitly take into account missing annotations. Improved performance could be achieved using a more sophisticated pseudo-labeling framework [46, 44, 18].

Acknowledgments

This work was conducted as part of the MINDS project of IRT Saint Exupéry. We would like to thank Michelle Aubrun, Ahmad Berjaoui, David Bertoin and Franck Mamalet for useful feedback and suggestions and Jérôme Mathieu for invaluable technical support.

References

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 549–565, 2016.

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[3] 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.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the IEEE European Conference on Computer Vision*, 2020.

[5] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 914–922, 2017.

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

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *Proceedings of the International Journal of Computer Vision*, pages 303–308, 2009.

[8] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems*, 2020.

[9] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 266–282, 2022.

[10] 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.

[11] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 752–768, 2020.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[13] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. In *Advances in Neural Information Processing Systems*, pages 16797–16807, 2020.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, pages 4904–4916, 2021.

[15] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, Zeng Yifu, Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, November 2022. URL https://doi.org/10.5281/zenodo.7347926.

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, pages 32–73, 2017.

[18] Hanjun Li, Xingjia Pan, Ke Yan, Fan Tang, and Wei-Shi Zheng. Siod: single instance annotated per category per image for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14197–14206, 2022.

[19] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=OJLaKwiXSbx.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[21] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[22] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the IEEE European Conference on Computer Vision*, 2020.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[24] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=MJIve1zgR_.

[25] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022.

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[27] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[28] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, 1997.

[29] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2641–2649, 2015.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[32] 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.

[33] 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 European Conference on Computer Vision*, pages 618–626, 2017.

[34] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018.

[35] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022.

[36] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *Proceedings of the British Machine Vision Conference*, 2021.

[37] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. In *arXiv:2005.04757*, 2020.

[38] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *Proceedings of the International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SygXPaEYvH`.

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019.

[40] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 176–191, 2018.

[41] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. In *Proceedings of the International Journal of Computer Vision*, pages 154–171, 2013.

[42] Huy V Vo, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Jean Ponce. Active learning strategies for weakly-supervised object detection. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 211–230, 2022.

[43] Keze Wang, Xiaopeng Yan, Dongyu Zhang, Lei Zhang, and Liang Lin. Towards human-machine cooperation: Self-supervised sample mining for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1605–1613, 2018.

[44] Tiancai Wang, Tong Yang, Jiale Cao, and Xiangyu Zhang. Co-mining: Self-supervised learning for sparsely annotated object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[45] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022.

[46] Mengmeng Xu, Yancheng Bai, Bernard Ghanem, Boxiao Liu, Yan Gao, N Guo, X Ye, F Wan, H You, D Fan, et al. Missing labels in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, page 5, 2019.

[47] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, and Yongfeng Huang. Image captioning with object detection and localization. In *Proceedings of the International Conference on Image and Graphics*, pages 109–118, 2017.

[48] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9686–9695, 2019.

[49] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, pages 67–78, 2014.

[50] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[51] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[52] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[53] 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.

[54] Han Zhang, Fangyi Chen, Zhiqiang Shen, Qiqi Hao, Chenchen Zhu, and Marios Savvides. Solving missing-annotation object detection with background recalibration loss. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1888–1892, 2020.

[55] Hanwang Zhang, Xindi Shang, Wenzhuo Yang, Huan Xu, Huanbo Luan, and Tat-Seng Chua. Online collaborative learning for open-vocabulary visual classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2809–2817, 2016.

[56] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 928–936, 2018.

[57] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[58] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 391–405, 2014.