Boosting Weakly Supervised Object Detection using Fusion and Priors from Hallucinated Depth

Cagri Gungor[1] and Adriana Kovashka[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science
University of Pittsburgh
cagri.gungor@pitt.edu, kovashka@cs.pitt.edu
https://cagrigungor.github.io/WSOD-AMPLIFIER/

Abstract

Despite recent attention to depth for various tasks, it is still an unexplored modality for weakly-supervised object detection (WSOD). We propose an amplifier method for enhancing the performance of WSOD by integrating depth information. Our approach can be applied to different WSOD methods based on multiple-instance learning, without necessitating additional annotations or inducing large computational cost. Our proposed method employs monocular depth estimation to obtain hallucinated depth information, which is then incorporated into a Siamese WSOD network using contrastive loss and fusion. By analyzing the relationship between language context and depth, we calculate depth priors to identify the bounding box proposals that may contain an object of interest. These depth priors are then utilized to update the list of pseudo ground-truth boxes, or adjust the confidence of per-box predictions. We evaluate our proposed method on three datasets (COCO, PASCAL VOC, and Conceptual Captions) by implementing it on top of two state-of-the-art WSOD methods, and we demonstrate a substantial enhancement in performance.

1. Introduction

Weakly-supervised object detection (WSOD) is a challenging task since it is unclear which instances have the label that was provided at the image level. Traditional methods only use appearance information in RGB images. However, appearance information is insufficient to localize objects in complex, cluttered environments. On the other hand, humans are capable of finding useful information in complex environments because they rely on object function, not just appearance. For example, they might reason about which objects are within reach, which can be captured with depth from stereo vision [2]. The depth modality provides additional cues about the spatial relationships and geometrical structure of objects in a scene and is invariant to appearance variations (e.g. in texture), making it complementary to the RGB modality. However, weakly-supervised object detection methods do not use depth information.

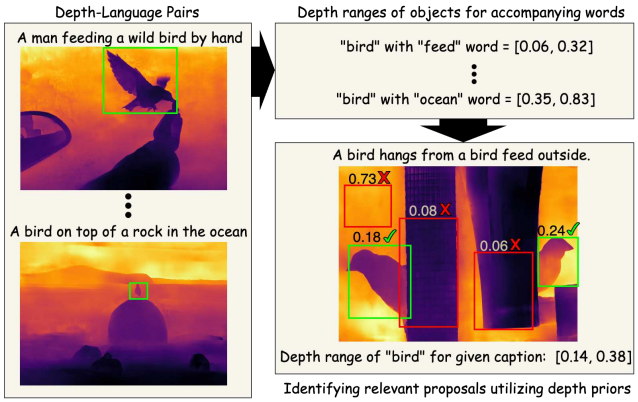

Figure 1. Object from the same category may be at different depth depending on the context/setting. We use captions to capture context-conditioned depth ranges for each object class and co-occurring word: a bird may be closer when co-occurring with the word "feed" than the word "ocean". We use these ranges to spot relevant proposals that may contain target objects, and prune irrelevant ones, in weakly supervised object detection training.

We equip WSOD methods with the ability to reason about functional information (depth). Importantly, our method does so without requiring additional annotations or suffering significant computational costs. We propose an amplifier method that can enhance the performance of different weakly supervised object detection methods based on multiple-instance learning. Since traditional WSOD datasets do not contain ground-truth depth information, the proposed method utilizes hallucinated (predicted) depth information obtained through a monocular depth estimation technique. During training, the method incorporates depth information to improve representation learning and to prune

or down-weight predictions at the box level, which leads to improved object detection performance during inference.

First, depth can directly be used as a feature to aid representation learning, or to produce predictions which can be fused with those computed from RGB. While simple, this technique has not been used for WSOD, and we show that it is very effective: it boosts the performance of appearance-only methods by up to 2.6 mAP@0.5 (11% relative gain).

Further, depth can provide strong priors about which of the bounding box proposals in the noisy WSOD setting contain an object of interest. We examine the rough depth at which objects of particular categories occur, by computing the depth range of an object using the predictions of a WSOD network. To make this range more precise, we examine the relationship between language context in captions and depth, by keeping track of depth range statistics conditioned on co-mentioned objects. The use of captions allows us to cope with the variable depth at which an object may occur depending on the context, as shown in Fig. 1. We then use this range to prune the pseudo ground-truth bounding boxes used to iteratively update weakly-supervised detection methods, or to down-weight predictions at the box level. This approach boosts WSOD performance further for a total up to 14% mAP@0.5 gain.

Our method is simple and can boost multiple WSOD methods that rely on iterative improvement. We test it using two state-of-art WSOD baselines, MIST [30] and SoS-WSOD [33], on COCO and PASCAL VOC. Inspired by recent work that trains object detection methods with language supervision [11, 36, 41], we further test our method in a setting where labels at the image level are not ground-truth but estimated. In this setting, our method boosts the basic WSOD performance even more, by 18% when labels for training are extracted from COCO, and 63% when they are extracted from Conceptual Captions.

To summarize, our contributions are: (1) We examine for the first time the use of depth in weakly-supervised object detection. (2) In addition to depth fusion, we propose a technique specific to WSOD, which estimates depth priors with the help of language, and uses them to refine pseudo boxes and box predictions. (3) We show large performance gains in a large variety of settings, with the biggest boost from depth refinement when supervision is least expensive.

2. Related Work

Weakly-supervised object detection (WSOD) is the task of learning to detect the location and type of objects given only image-level labels during training. The multi-instance learning (MIL) framework is commonly utilized in WSOD methods such as WSDDN [1]. OICR [35] improved upon this by proposing pseudo-ground-truth mining and an online instance refinement, which was further refined by proposal clustering [34]. C-MIL [37] and MIST [30] introduced modifications to the MIL loss and pseudo-ground-truth mining, respectively. SoS-WSOD [33] proposed a method that produces pseudo boxes for FSOD and splits noisy data for semi-supervised object detection. Additionally, there have been efforts to bypass the need for image-level labels by utilizing noisy labels extracted from caption or subtitle data [4, 11, 36, 38, 41]. Additionally, [12] leverages audio to improve WSOD performance and reduce noise from text-based label extraction. In contrast to these works, our method leverages depth information as an additional modality, leading to improved performance in WSOD and a reduction of the noise in labels extracted from text.

RGB-D detection. The integration of RGB and depth information to derive complementary features has been previously studied for fully-supervised indoor analysis [19, 22, 40, 42, 43] and object detection [8, 9, 15–17, 20, 21]. The strategies for merging the two modalities can be classified into three groups, depending on the point in the processing pipeline where the fusion occurs: early fusion [6, 27], middle fusion [3, 9, 10, 44], and late fusion [14, 26]. Early fusion techniques involve combining the RGB and depth images into a single four-channel matrix at the earliest stage of the process. Middle fusion provides a balance between early and late fusion by utilizing CNNs for both feature extraction and subsequent merging. In late fusion, individual saliency prediction maps are produced from the RGB and depth channels to be combined through post-processing operations. In contrast to the majority of aforementioned methods, which use separate networks to extract features from RGB and depth images, several studies [9, 10, 24, 32] employ Siamese networks to learn hierarchical features from both RGB and depth inputs by utilizing shared parameters. However, *we are the first to leverage depth data in weakly-supervised object detection.* Our approach is not specific to a particular method, as it can be applied to different MIL-based WSOD methods to improve their performance without incurring any extra annotation expenses and with minimal computational overhead during training. Although the depth modality is not used during the inference stage, incorporating it during training enhances the performance of the inference.

Monocular depth estimation involves predicting the depth map of a scene from a single RGB image [25, 28, 29, 39]. We utilize the method in [25] to estimate depth on the training set due to its strong performance. This estimated ("hallucinated") depth information is utilized to improve the performance of weakly supervised object detection.

3. Approach

We propose an amplifier approach that incorporates a depth modality to improve the effectiveness of WSOD methods. Our method can be used with different MIL-based WSOD methods to boost their performance by in-

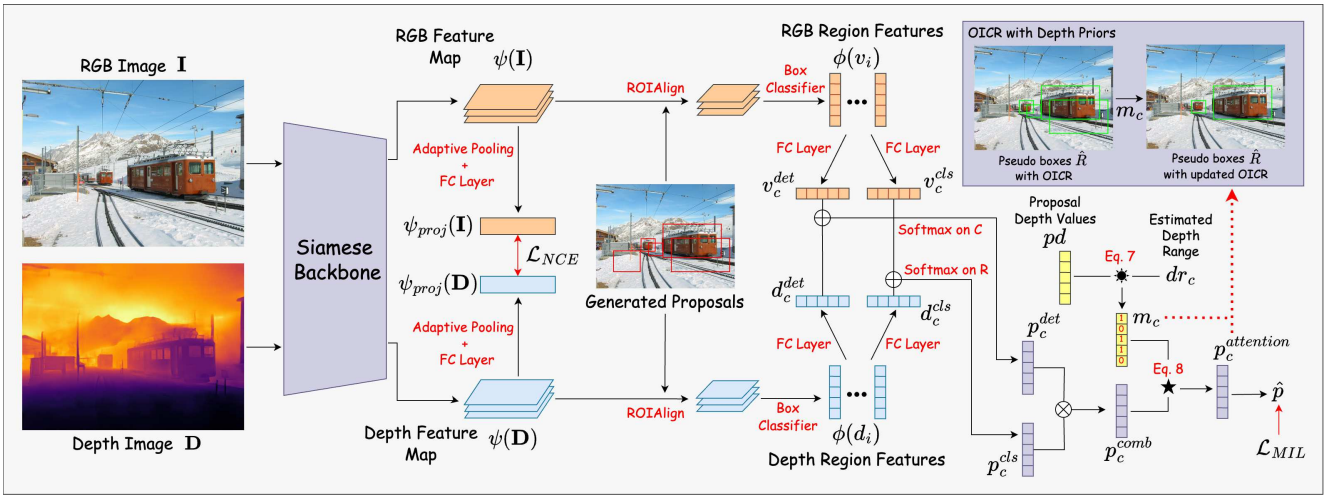

Figure 2. This figure illustrates the design of our proposed amplifier technique that takes advantage of depth information to enhance the performance of other weakly-supervised object detection methods. During inference, we only use the RGB branch (shown in orange).

curring little extra cost during training. It does not use the depth modality during inference to avoid any slow-downs and reliance on additional data (depth estimation or captions). The proposed approach comprises three main steps (Sec. 3.1, 3.2, 3.3, respectively). First, a Siamese network with a shared backbone is employed to improve representation learning through contrastive learning between RGB and depth features (referred to as SIAMESE-ONLY in the experiments). Second, we combine detection and classification scores obtained from both the RGB and depth modalities, which can be categorized as late fusion (FUSION). Third, we use captions and bounding box predictions of traditional WSOD to calculate depth priors. These depth priors are then used to improve the OICR-style [35] module in two WSOD methods (named DEPTH-OICR) and create attention with combined score probabilities (DEPTH-ATTENTION). Note that SIAMESE-ONLY is always applied, while FUSION, DEPTH-OICR and DEPTH-ATTENTION build on top of it, and can be used alone or combined.

3.1. The Siamese WSOD Network

WSOD. Following Bilen et al. [1], let $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$ denote an RGB image, and $y_c \in \{0, 1\}$ (where $c \in \{1, ..., C\}$ and C is the total number of object categories) be its corresponding ground-truth class labels. Let v_i, $i \in \{1, ..., R\}$ (where R is the number of proposals), denote the visual proposals in image $\mathbf{I}$. RoI pooling is applied and a fixed-length feature vector $\phi(v_i)$ extracted for each visual region. The proposal features $\phi(v_i)$ are fed into two parallel fully-connected layers to compute the visual detection score $v_{i,c}^{det} \in \mathbb{R}^1$ and classification score $v_{i,c}^{cls} \in \mathbb{R}^1$:

$$v_{i,c}^{det} = w_c^{det\intercal}\phi(v_i) + b_c^{det}, \quad v_{i,c}^{cls} = w_c^{cls\intercal}\phi(v_i) + b_c^{cls} \quad (1)$$

where w and b are weights and bias, respectively.

Estimating the depth images. To extract depth information from RGB images, we employ the monocular depth estimation technique by Mahdi et al. [25]. This enables us to use existing RGB-only object detection datasets without the need for additional annotations. Although the extracted depth images are initially grayscale, we use a color map to convert them to RGB images with three channels.

Siamese design. Our approach utilizes a Siamese network with contrastive learning to incorporate depth information in the weakly-supervised object detection network during training. This design allows us to use a backbone pre-trained with RGB images to extract features from both RGB and depth images, without adding extra complexity to the model's parameters. We enhance the representation learning of the backbone by defining contrastive loss between RGB and depth features similar to [24]. Utilizing a Siamese network provides the advantage of using only RGB images during inference similar to other WSOD methods. This ensures that our contribution does not introduce any additional overhead on the inference time.

With the help of a pre-trained backbone model, the feature map of RGB image $\psi(\mathbf{I})$ is extracted. Let $\mathbf{D} \in \mathbb{R}^{h \times w \times 3}$ denote a depth image associated with the RGB image $\mathbf{I}$ and let $\psi(\mathbf{D})$ be the feature map of the depth image $\mathbf{D}$ extracted by the Siamese backbone. The RGB feature map $\psi(\mathbf{I})$ and depth feature map $\psi(\mathbf{D})$ are fed into adaptive pooling and fully connected layers to obtain d-dimensional projected feature vectors $\psi_{proj}(\mathbf{I})$ and $\psi_{proj}(\mathbf{D})$. The only extra parameters we add to the traditional MIL-based WSOD network come from the fully connected layer for projection with 8 percent overhead (13M parameters for the projection layer, vs 154M total). If no late fusion is per-

formed in the experiments, we train as described in Sec. 3.2, but excluding the $d_{i,c}$ variables in Eq. 5.

Contrastive learning. We L2-normalize the RGB and depth feature vectors $\psi_{proj}(\mathbf{I})$ and $\psi_{proj}(\mathbf{D})$ vectors, and compute their cosine similarity:

$$S(\mathbf{I}, \mathbf{D}) = \langle \psi_{proj}(\mathbf{I}), \psi_{proj}(\mathbf{D}) \rangle / \rho \tag{2}$$

where ρ, is a learnable temperature parameter. We use noise contrastive estimation (NCE) [13] to define the contrastive learning by considering RGB image and depth image pairs $(\mathbf{I}, \mathbf{D}) \in \mathcal{B}$ where $\mathcal{B}$ is an RGB-depth pair batch. The first component of the NCE loss contrasts an RGB image with negative depth images to measure how closely the RGB image matches with its paired depth among others in the batch:

$$\mathcal{L}_{D \to I} = -\frac{1}{|\mathcal{B}|} \sum_{(\mathbf{I},\mathbf{D}) \in \mathcal{B}} log \frac{exp(S(\mathbf{I}, \mathbf{D}))}{exp(S(\mathbf{I}, \mathbf{D})) + \sum_{(\mathbf{I}', \mathbf{D}') \in \mathcal{B}} exp(S(\mathbf{I}, \mathbf{D}'))} \tag{3}$$

The second component of the NCE loss, $\mathcal{L}_{I \to D}$, is analogously defined to contrast a depth image with negative RGB image samples, and the two components are averaged:

$$\mathcal{L}_{NCE} = (\mathcal{L}_{D \to I} + \mathcal{L}_{I \to D})/2 \tag{4}$$

3.2. Late Fusion of the Modalities

The detection and classification scores computed from RGB and depth modalities are imbued with disparate and complementary details that jointly enrich our understanding of the target objects. Therefore, we combine these scores to amplify the performance of object detection.

As the depth images are derived from the RGB images, the spatial arrangement of the objects is equivalent in both modalities. Hence, we utilize the same visual region proposals for both RGB and depth modalities. Following the application of the RoI pooling layer and the Siamese box feature extractor to the depth feature map $\psi(\mathbf{D})$, we obtain the feature vector $\phi(d_i)$ for each depth region. Thereafter, we employ the approach presented in Eq. 1 to derive the depth detection score $d^{det}_{i,c} \in \mathbb{R}^1$ and the depth classification score $d^{cls}_{i,c} \in \mathbb{R}^1$. Subsequently, we fuse (sum) the scores from the RGB and depth modalities:

$$f^{det}_{i,c} = v^{det}_{i,c} + d^{det}_{i,c}, \quad f^{cls}_{i,c} = v^{cls}_{i,c} + d^{cls}_{i,c} \tag{5}$$

where $f^{det}_{i,c}$ and $f^{cls}_{i,c}$ are fusion detection and classification scores, respectively.

Following the WSDDN [1] architecture, these classification and detection scores are converted to probabilities such that $p^{cls}_{i,c}$ is the probability that class c is in present proposal f_i, and $p^{det}_{i,c}$ is the probability that f_i is important for predicting image-level label y_c.

$$p^{det}_{i,c} = \frac{exp(f^{det}_{i,c})}{\sum_{k=1}^{R} exp(f^{det}_{k,c})}, \quad p^{cls}_{i,c} = \frac{exp(f^{cls}_{i,c})}{\sum_{k=1}^{C} exp(f^{cls}_{i,k})} \tag{6}$$

We element-wise multiply the classification and detection scores to obtain the combined score $p^{comb}_{i,c}$:

$$p^{comb}_{i,c} = p^{det}_{i,c} p^{cls}_{i,c} \tag{7}$$

Finally, image-level predictions $\hat{p}_c$ are computed as follows, where greater values of $\hat{p}_c \in [0, 1]$ mean a higher likelihood that c is present in the image.

$$\hat{p}_c = \sigma \left(\sum_{i=1}^{R} p^{comb}_{i,c} \right) \tag{8}$$

Assuming the label $y_c = 1$ if and only if class c is present, the classification loss used for training the model is defined as follows. Since no region-level labels are provided, we must derive region-level scores indirectly, by optimizing this loss.

$$\mathcal{L}_{mil} = -\sum_{c=1}^{C} [y_c \log \hat{p}_c + (1 - y_c) \log(1 - \hat{p}_c)] \tag{9}$$

3.3. Depth Priors

We utilize the baseline WSOD methods, which we aim to improve, to generate bounding box object predictions in the training set. Further, we leverage both the generated bounding box predictions and associated captions to extract knowledge about the relative depths of objects. We note that our proposed methodology adheres to the WSOD setting, deriving benefits from the predicted bounding boxes, as opposed to ground truth bounding box annotations to calculate depth priors. We subsequently exploit these depth priors to guide the identification of the relevant visual regions that may contain the target objects. Further, we show that even though we estimate the depth priors from COCO, they generalize to Conceptual Captions (Table 2).

We use the notation $pd_i \in [0, 1]$, $i \in 1, ..., R$, where R is the number of pre-computed region proposals for depth image $\mathbf{D}$, to represent the average depth value in the i-th region proposal. Each region proposal contains pixels with values ranging from 0 to 1, which correspond to the smallest and largest depth values, respectively.

We employ bounding box predictions B to approximate the depth value of objects using the caption that describes the image in which the objects are present. We also use co-occurring captions to capture the context in which an object occurs, and condition depth priors on this context which varies across images. Let C be the set of object categories, W be the set of distinct words in the vocabulary that includes every word in the captions, and B be the set of predicted bounding boxes. Let $d_{c,w,b} \in \{[0, 1], \varnothing\}$ denote the depth value for object $c \in C$, word $w \in W$ and box $b \in B$, which is calculated by averaging the depth values in the pixels of b similar to the calculation of pd_i. As an example, $d_{\text{bird,ocean},b}$ represents the depth value of the "bird" object

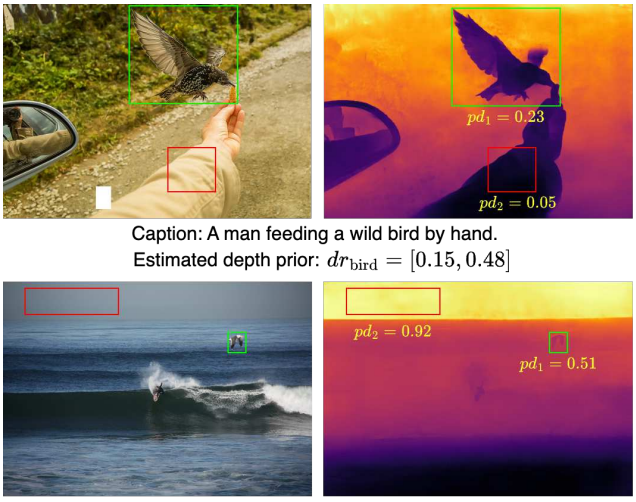

Figure 3. The figure displays a row of images that are accompanied by their respective depth and caption data, as well as proposal depth value of different regions and estimated depth prior range.

box b of a depth image that has a caption that includes the "ocean" word. In the absence of "ocean" in the caption or when annotation b does not correspond to the "bird" object, the depth value $d_{\text{bird,ocean},b}$ is set to null $\varnothing$. Further, $d_{c,w}$ represents a set of depth values calculated by averaging $d_{c,w,b}$ over predicted boxes $b \in B$, excluding $\varnothing$ values. The depth range $r_{c,w} = [mean - std, mean + std]$ for class c and word w is obtained by utilizing the mean and standard deviation (std) of this set of depth values in $d_{c,w}$. Once these depth ranges $r_{c,w}$ are computed, they can be applied to estimate an allowable depth range for a class c in a new image, without any boxes on that image.

For any new depth image $\mathbf{D}$, the range of estimated depth priors for an object c is dr_c:

$$dr_c = \frac{\sum_{s \in S} r_{c,s}}{|S|} \tag{10}$$

where S denotes the set of words in the caption corresponding to $\mathbf{D}$. Thus the depth at which we expect to find objects in a particular image varies depending on the context provided by words in the corresponding caption. We **only require captions at training time.**

We utilize the estimated depth prior range dr_c to identify potentially important regions in pd for each class. We define a depth mask indicator variable $m_{i,c} \in 0, 1$ for each region $i \in R$ and class $c \in C$, which indicates the likelihood of a particular region in an image containing an object of a certain class. The computation of this variable is as follows:

$$m_{i,c} = \begin{cases} 1, & \text{if } pd_i \in dr_c \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

If the proposal depth value pd_i falls within the estimated depth prior range dr_c for class c, it is considered as a relevant region for that class, and the corresponding mask variable $m_{i,c}$ is set to 1; otherwise, it is set to 0. Subsequently, we utilize the mask variable $m_{i,c}$ in combination with our end-to-end network to improve its performance.

As an example, Fig. 3 presents two images featuring a "bird" object, with different depths. The estimated depth prior ranges dr_{bird} are calculated using Eq. 10 for each image based on the words in the caption. The caption of the first image includes "feeding" and "hand" which suggest the "bird" is likely to have a smaller depth, while the caption of the second image includes "flying" and "ocean" that suggest the "bird" is likely to have a bigger depth. The regions on the images having a proposal depth value of pd_1 are in the estimated depth prior range dr_{bird}; we observe that they truly include the "bird" object. The range allows us to rule out regions with values pd_2, which do not contain "bird".

Alternative method for estimating depth priors. As an alternative, we use only bounding box predictions (without captions) to obtain depth priors. Let $d_{c,b} \in [0, 1]$ denote the depth value for object $c \in C$ and box $b \in B$. Further, let d_c represent a set of depth values for each c. The depth range $r_c = [mean - std, mean + std]$ is obtained by utilizing the mean and standard deviation (std) of this set of depth values in d_c. Then we set $dr_c = r_c$ as we do not use caption information (compared to Eq. 10); dr_c is used in Eq. 11.

3.3.1 Depth Priors: Updated OICR

Algorithm 1 OICR Mining with Depth Priors

Input: Proposals R, Depth Mask Indicator Variable m
Output: Pseudo boxes $\hat{R}$

1: $\hat{R} = \varnothing$
2: **for all** $c = 1 : C$ **do**
3: **for all** $i = 1 : |R|$ **do**
4: $\hat{R}_c = \hat{R}_c \cup R_i$ **if** $m_{i,c} = 1$
5: **return** $\hat{R}$

Online Instance Classifier Refinement (OICR) [35] is a weakly supervised object detection algorithm that iteratively refines object proposals. Recent studies [30, 33, 34] have highlighted the importance of more effective proposal mining strategies for achieving better recall and precision of objects in WSOD detectors. We propose an algorithm that incorporates the depth priors during the proposal mining provided in Alg. 1. As our proposed method aims to enhance MIL-based WSOD methods, we utilize our algorithm in conjunction with recent OICR-style/self-training/mining strategies, subject to the depth prior condition specified in the fourth line of Alg. 1. After using the depth prior condition, OICR-style mining selects fewer but more relevant

proposals so our contribution increases mining precision.[1]

3.3.2 Depth Priors: Attention

The depth mask variable $m_{i,c}$ indicates the potentially important proposal regions for each class. We use this variable to employ an attention mechanism with combined score probabilities $p_{i,c}^{comb}$ provided in Eq. 7 as follows:

$$p_{i,c}^{comb} = p_{i,c}^{comb} * 0.5, \text{if } m_{i,c} = 0 \tag{12}$$

This mechanism reduces the probability of a region for class c by half if the region is determined as less likely to be important by $m_{i,c}$. These scores are then used in Eq. 8.

4. Experiments

We test our method on top of two weakly-supervised detection techniques, and verify the contributions of the constituents of our approach:

- Siamese WSOD Network (SIAMESE-ONLY, Sec. 3.1);
- Late Fusion of the Modalities (FUSION, Sec. 3.2) which combines classification/detection from RGB and depth, and builds on top of the Siamese WSOD Network (Sec. 3.1);
- Depth Priors are utilized to enhance the OICR-style module (DEPTH-OICR, Sec. 3.3.1) and construct attention (DEPTH-ATTENTION, Sec. 3.3.2) with visual-only score probabilities, both building upon the Siamese WSOD Network (Sec. 3.1);
- Finally, we use all components of our method (WSOD-AMPLIFIER, SEC. 3.1, 3.2, 3.3.1, 3.3.2).
- DEPTH-OICR-ALT and DEPTH-ATTENTION-ALT estimate depth priors without captions.
- WSOD-AMPLIFIER-INF fuses RGB and depth at inference time, unlike our proposed method.

4.1. Experimental Setup

PASCAL Visual Object Classes 2007 (VOC-07) [5] contains 20 classes. For training, we use 2501 images from the train set and 2510 images from the validation set. We evaluate using 4952 images from the test set.

Common Objects in Context (COCO) [23] consists of 80 classes. We utilize approximately 118k images from the train set and use the labels provided at the image level. Additionally, to test how well our method works when labels are obtained from noisy language supervision (in captions), we train our models using labels obtained through an exact match (EM) method following [36], also referred to as substring matching in [7]. Due to the unavailability of any labels for around 15k images extracted from captions, we excluded them from the training set and use 103k images. We evaluate using 5k images from the validation set.

Conceptual Captions (CC) [31] is a large-scale image captioning dataset containing over 3 million images annotated with only captions. We use around 30k images and their corresponding captions and the labels are extracted for the 80 COCO dataset classes using an exact match method from the captions. During the evaluation, we used 5k images from the COCO validation set.

Domain shift datasets. In the supplementary, we also evaluate our method in a domain shift setting [18], using three datasets. Clipart1k has the same 20 classes as VOC with 1,000 images, while Watercolor2k and Comic2k share 6 classes with VOC and have 2,000 images each.

Evaluation protocols. We utilize mean Average Precision (mAP) considering various IoU thresholds as the common evaluation metric for COCO and VOC datasets. Additionally, we report mAP for objects of different sizes during COCO evaluation and we report the results of Correct Localization (CorLoc) for VOC evaluation.

Implementation details. We employ the official PyTorch implementations of SoS-WSOD [33] and MIST [30] methods to apply our amplifier technique. SoS-WSOD uses four images per GPU as two augmented images and their flipped versions with a total of 4 GPUs, whereas MIST uses only one image per GPU with a total of 8 GPUs. However, we use one image per GPU for SoS-WSOD due to VRAM limitation in our GPUs, as we also utilized depth images for each corresponding RGB image. Therefore, the baseline results of SoS-WSOD reported in Table 1 are slightly lower than those reported in the original paper. Moreover, we solely use the first stage of SoS-WSOD since it includes the MIL-based WSOD module which is convenient to implement our method on top of. The other settings are kept the same as the official implementations with the VGG16 backbone. The inference is done on the training set by using baseline MIST and SoS-WSOD methods to obtain box predictions having confidence scores higher than 0.5. These box predictions are then used to calculate depth priors. Furthermore, we utilize the same depth range $r_{c,w}$ from the COCO annotations for the WSOD-AMPLIFIER method on the Conceptual Captions dataset.

4.2. Comparing our amplifier to state of the art

We evaluate our proposed methods, FUSION and WSOD-AMPLIFIER, using two state-of-the-art WSOD approaches, SoS-WSOD [33] and MIST [30], and the COCO and VOC-07 datasets. The performance of our proposed methods are compared with the baseline methods in Table 1. When our WSOD-AMPLIFIER method is applied to MIST, it improves the baseline performance by 17% in $mAP_{50:95}$ (relative gain, 13.8/11.8-1) and 14% in mAP_{50}. Similarly, when our WSOD-AMPLIFIER method is applied to SoS-

[1] In early experiments, we verified our method's gains persist if the baseline drops the lowest-scoring pseudo boxes without using depth.

Methods on COCO	Avg. Precision, IoU			Avg. Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
MIST [30]	11.8	24.3	10.7	3.6	13.2	18.9
+ FUSION	13.5	26.9	12.4	4.0	14.7	21.6
+ WSOD-AMPLIFIER	**13.8**	**27.8**	**12.5**	**4.6**	**14.8**	**22.6**
+ WSOD-AMPLIFIER-INF	13.1	27.5	11.9	4.3	14.3	22.2
SOS-WSOD [33]	10.2	21.5	8.6	**2.5**	10.6	17.7
+ FUSION	10.3	21.6	8.9	2.3	10.8	18.4
+ WSOD-AMPLIFIER	**10.5**	**21.8**	**9.1**	**2.5**	**11.1**	**18.7**
Methods on VOC-07	Avg. Precision, IoU			CorLoc		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
SOS-WSOD [33]	24.8	52.2	20.4	38.7	71.7	36.9
+ FUSION	**26.0**	**53.1**	**22.3**	**39.6**	**72.1**	**38.5**

Table 1. This table compares the performance enhancement of our methods, to their baseline results SoS-WSOD [33] and MIST [30], on COCO and VOC-07. The best performer per column is in **bold**.

Methods on COCO	Avg. Precision, IoU			Avg. Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
MIST [30] w/ GT	9.7	21.1	8.0	3.0	10.4	15.1
MIST [30] w/ EM	8.5	17.9	7.3	3.0	9.4	14.9
+ SIAMESE-ONLY	8.8	18.7	7.3	2.9	9.6	15.4
+ DEPTH-OICR	9.0	19.4	7.3	3.1	9.6	15.9
+ DEPTH-ATTENTION	9.1	19.0	7.9	3.0	9.5	16.2
+ FUSION	9.9	20.4	8.5	3.0	10.1	17.1
+ WSOD-AMPLIFIER	**10.2**	**21.0**	**8.5**	**3.3**	**10.3**	**17.5**
+ DEPTH-OICR-ALT	8.9	19.0	7.3	3.0	9.6	15.6
+ DEPTH-ATTENTION-ALT	9.0	18.8	7.7	3.0	9.5	16.0
Methods on CC	Avg. Precision, IoU			Avg. Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
MIST [30] w/ EM	1.7	3.8	1.4	0.3	1.7	3.4
+ FUSION	2.0	4.1	1.7	0.3	1.9	4.0
+ DEPTH-OICR	2.4	5.6	2.0	0.3	2.2	5.1
+ WSOD-AMPLIFIER	**2.6**	**6.3**	**2.1**	**0.4**	**2.8**	**5.7**

Table 2. This table introduces the effect of each component of our method implemented on MIST [30] with exact match (EM) labels on COCO (top) and Conceptual Captions (CC) (bottom). The best performer per column is in **bold**. On top, all proposed methods that outperform the SIAMESE-ONLY are underlined.

WSOD, it improves the baseline performance by 2% in $mAP_{50:95}$, 1.5% in mAP_{50}, and 6% in mAP_{75}. As the VOC-07 dataset does not have captions, we are only able to apply the SIAMESE-ONLY and FUSION methods on SoS-WSOD but not the DEPTH-OICR and DEPTH-ATTENTION. On this dataset, our improvements outperformed the baseline SoS-WSOD by 5% in $mAP_{50:95}$, 2% in mAP_{50}, and 9% in mAP_{75}. WSOD-AMPLIFIER-INF performs worse than WSOD-AMPLIFIER. We argue depth is useful as a soft guide to balance region information during MIL training, but less so when directly used in the strict detection setting.

4.3. Ablation studies and visualization

Experiments with labels from captions. Several attempts [7, 36, 38] have been made to eliminate the requirement for image-level labels by leveraging noisy label information obtained from captions or subtitles. Although it is cost-effective to use text information for label extraction, it results in a decrease in the performance of weakly supervised object detection. [36] propose a text classifier approach to extract labels more effectively than the simple exact match (EM) and reduce the noise between text and ground truth (GT) labels. In contrast to previous studies, our research employs the depth modality to reduce the noise in labels extracted from captions. Our approach improves the model's detection capability and employs captions during the calculation of depth priors. We conducted experiments with MIST [30] using both GT and EM labels and observed that, as expected, training with GT labels leads to significantly better performance than training with EM labels in Table 2 due to the noise in labels extracted from captions. However, our proposed WSOD-AMPLIFIER method applied on MIST with EM labels surpasses the baseline and MIST with GT labels. These findings demonstrate that our method *effectively reduces noise* and enables the model trained with EM labels to achieve better performance than those trained with GT labels. It is worth noting that the text classifier approach proposed by [36] also performs better than EM-labeled training data, but falls short of the performance achieved by GT-labeled data.

Results on noisy datasets. We also extract labels from captions on the Conceptual Captions dataset, which lacks labels at the image level. We observe that our WSOD-AMPLIFIER boosts results by an impressive 63% relative gain using mAP_{50}. Conceptual Captions is a noisier dataset than COCO, since captions were not collected through crowdsourcing, but were crawled as alt-text for web search results. Thus, it is noteworthy that *the benefit of our approach becomes more pronounced as the cost of supervision decreases, and the noise in the supervision increases.*

Analysing the components of our approach. To understand the impact of each component of our approach on the overall performance, we conducted experiments with MIST [30] using EM labels as a baseline and applied each component of our method on top of the baseline in Table 2. Our SIAMESE-ONLY method, which incorporates the depth modality in the Siamese network using contrastive learning, improves feature extraction and results in a 4% increase in $mAP_{50:95}$ and mAP_{50}. Our DEPTH-OICR method, which utilizes depth priors in the OICR module to improve the mining strategy, increases $mAP_{50:95}$ and mAP_{50} over SIAMESE-ONLY by $6-8\%$ on COCO and $42-47\%$ on Conceptual Captions (CC). Our DEPTH-ATTENTION method, which incorporates depth priors to use potentially important regions in an attention mechanism

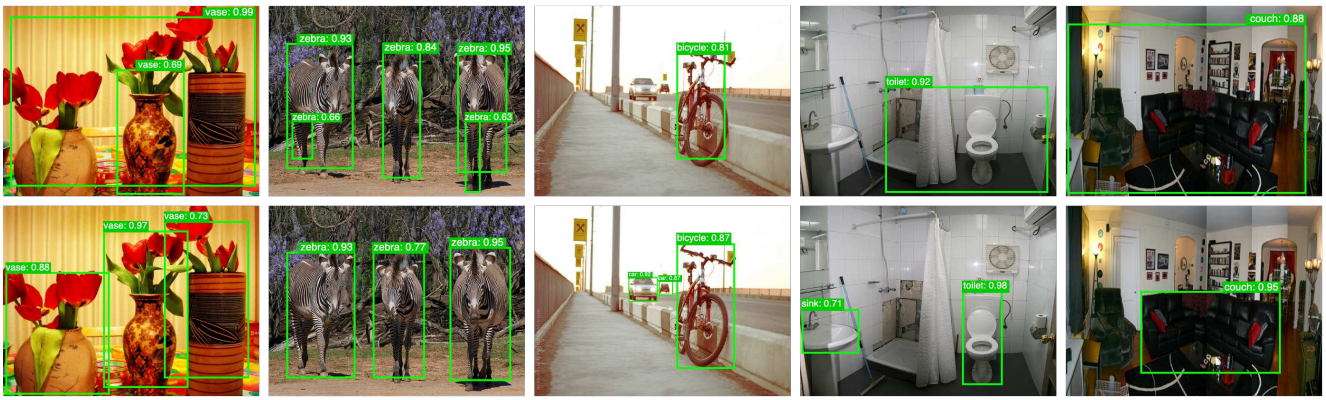

Figure 4. Qualitative comparison of MIST [30] (top) and our proposed WSOD-AMPLIFIER method (bottom), on COCO val. The ground-truth objects are vase, zebra, bicycle and car, sink and toilet, couch in this order. Confidence scores and names of objects shown.

with combined score probabilities, increases $AP_{50:95}$ and mAP_{75} by $6-7\%$. Our FUSION method, which combines RGB and depth image scores, improves by $14-16\%$ on COCO and $7-21\%$ on CC. Comparing FUSION and DEPTH-OICR, the bigger gain using mAP_{50} is achieved by FUSION on COCO, and DEPTH-OICR on CC. *Thus, the benefit of our WSOD-specific method component increases as the noise in the dataset increases, which is appealing due to its real-world applicability.* Finally, our WSOD-AMPLIFIER method, which includes all components of our approach, achieves the highest performance increase over MIST w/ EM baseline, with *improvements in all* mAP *metrics by* $16-20\%$ *on COCO and* $53-65\%$ *on CC.*

Alternative depth priors. At the bottom of Table 2 (COCO), we see that DEPTH-OICR-ALT and DEPTH-ATTENTION-ALT, derived from the alternative approach, yield superior results compared to SIAMESE-ONLY. However, the DEPTH-OICR and DEPTH-ATTENTION methods, derived from our full (caption-based) approach, outperform both DEPTH-OICR-ALT and DEPTH-ATTENTION-ALT. Note that the depth range r_c remains consistent across all images in this alternative method. Conversely, in the proposed approach, the depth range dr_c is computed individually for each image by taking into account the corresponding caption, as visualized in Fig. 3. As a result, the proposed approach demonstrates enhanced capacity in modeling depth priors through the utilization of captions.

Generalization of depth priors. Even though on CC we use the depth priors calculated from COCO, our proposed method exhibits a more substantial enhancement in CC performance compared to COCO, achieving an improvement of 63% mAP_{50} over the MIST baseline (50% improvement from DEPTH-OICR alone). Thus, our DEPTH-OICR demonstrates generalization, as it has a higher impact than FUSION on CC (without recomputing priors), in contrast to COCO. Given the recent interest in learning from vision-language data, our approach has the potential to be highly impactful. Further, we compared the priors estimated from different datasets, and found them to be similar. In particular, 82.3% of PASCAL objects fit within the $[mean - stdev, mean + stdev]$ range computed from COCO, and 84.4% when the range is computed on PASCAL itself; the cross-domain gap in the range is small.

Qualitative analysis. We visualize the object detection performance of our proposed WSOD-AMPLIFIER compared to MIST [30] in Fig. 4. The confidence scores are calculated using visual detection v^{det} and classification scores v^{cls}. We show boxes with scores higher than 0.5. In the first image, the baseline struggles to accurately identify multiple instances of the same "vase" objects, instead grouping them together in a single box. Our method overcomes this challenge, precisely detecting each individual "vase". In the second image, the baseline faces the problem of part domination due to some discriminative parts of a "zebra". Our method overcomes this issue by utilizing depth modality during training, which emphasizes the geometric variations of objects, while comparatively ignoring the complex background. In other images, unlike our method, the baseline misses objects entirely, or produces large and imprecise bounding boxes. Moreover, the boxes detected by our method tend to have higher prediction scores.

Conclusion. We show depth boosts weakly-supervised object detection methods, tested on SoS-WSOD and MIST, without extra annotation or costly computation. Our Siamese WSOD network efficiently incorporates RGB and depth with contrastive learning and fusion. Using the relation of language and depth, depth priors estimate the bounding box proposals that may contain an object of interest.

Acknowledgement: This work was supported by a National Science Foundation Award No. 2046853, and a University of Pittsburgh Momentum Funds award.

References

[1] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016. 2, 3, 4

[2] P Cammack and JM Harris. Depth perception in disparity-defined objects: finding the balance between averaging and segregation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1697):20150258, 2016. 1

[3] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3051–3060, 2018. 2

[4] Kai Chen, Hang Song, Chen Change Loy, and Dahua Lin. Discover and learn new objects from documentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–308, 2009. 6

[6] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2075–2089, 2020. 2

[7] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning (ICML)*, pages 6216–6234. PMLR, 2022. 6, 7

[8] Guang Feng, Jinyu Meng, Lihe Zhang, and Huchuan Lu. Encoder deep interleaved network with multi-scale aggregation for rgb-d salient object detection. *Pattern Recognition*, 128:108666, 2022. 2

[9] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3052–3062, 2020. 2

[10] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for rgb-d salient object detection and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(9):5541–5559, 2021. 2

[11] 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 *European Conference on Computer Vision (ECCV)*, 2022. 2

[12] Cagri Gungor and Adriana Kovashka. Complementary cues from audio help combat noise in weakly-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2185–2194, 2023. 2

[13] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 4

[14] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, 48(11):3171–3183, 2017. 2

[15] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, 2016. 2

[16] Haiwen Huang, Andreas Geiger, and Dan Zhang. Good: Exploring geometric cues for detecting objects in an open world. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[17] Tanveer Hussain, Abbas Anwar, Saeed Anwar, Lars Petersson, and Sung Wook Baik. Pyramidal attention for saliency detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2877–2887. IEEE, 2022. 2

[18] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5009, 2018. 6

[19] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson WH Lau, and Thomas S Huang. Geometry-aware distillation for indoor semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2869–2878, 2019. 2

[20] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 630–647. Springer, 2022. 2

[21] Jingjing Li, Wei Ji, Qi Bi, Cheng Yan, Miao Zhang, Yongri Piao, Huchuan Lu, et al. Joint semantic mining for weakly supervised rgb-d salient object detection. *Advances in Neural Information Processing Systems*, 34:11945–11959, 2021. 2

[22] Yabei Li, Zhang Zhang, Yanhua Cheng, Liang Wang, and Tieniu Tan. Mapnet: Multi-modal attentive pooling network for rgb-d indoor scene classification. *Pattern Recognition*, 90:436–449, 2019. 2

[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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[24] Johannes Meyer, Andreas Eitel, Thomas Brox, and Wolfram Burgard. Improving unimodal object recognition with multimodal contrastive learning. In *2020 IEEE/RSJ International*

Conference on Intelligent Robots and Systems (IROS), pages 5656–5663. IEEE, 2020. 2, 3

[25] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9685–9694, 2021. 2, 3

[26] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7254–7263, 2019. 2

[27] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017. 2

[28] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 2

[29] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(3):1623–1637, 2020. 2

[30] 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 (CVPR)*, 2020. 2, 5, 6, 7, 8

[31] 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 (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[32] Hwanjun Song, Eunyoung Kim, Varun Jampan, Deqing Sun, Jae-Gil Lee, and Ming-Hsuan Yang. Exploiting scene depth for object detection with multimodal transformers. In *32nd British Machine Vision Conference (BMVC)*, pages 1–14. British Machine Vision Association (BMVA), 2021. 2

[33] Lin Sui, Chen-Lin Zhang, and Jianxin Wu. Salvage of supervision in weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14227–14236, 2022. 2, 5, 6, 7

[34] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(1):176–191, 2018. 2, 5

[35] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 5

[36] Mesut Erhan Unal, Keren Ye, Mingda Zhang, Christopher Thomas, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Learning to overcome noise in weak caption supervision for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2, 6, 7

[37] 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 (CVPR)*, 2019. 2

[38] 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 (CVPR)*, pages 9686–9695, 2019. 2, 7

[39] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 204–213, 2021. 2

[40] Xiaowen Ying and Mooi Choo Chuah. Uctnet: Uncertainty-aware cross-modal transformer network for indoor rgb-d semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 20–37. Springer, 2022. 2

[41] 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 (CVPR)*, pages 14393–14402, 2021. 2

[42] Zhijie Zhang, Yan Liu, Junjie Chen, Li Niu, and Liqing Zhang. Depth privileged object detection in indoor scenes via deformation hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3456–3464, 2021. 2

[43] Feng Zhou, Yu-Kun Lai, Paul L Rosin, Fengquan Zhang, and Yong Hu. Scale-aware network with modality-awareness for rgb-d indoor semantic segmentation. *Neurocomputing*, 492:464–473, 2022. 2

[44] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. Pdnet: Prior-model guided depth-enhanced network for salient object detection. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 199–204. IEEE, 2019. 2