Open-Vocabulary Object Detection Using Captions

Alireza Zareian[1,2],   Kevin Dela Rosa[1],   Derek Hao Hu[1],   Shih-Fu Chang[2]
[1] Snap Inc., Seattle, WA   [2] Columbia University, New York, NY
{azareian, kevin.delarosa, hao.hu}@snap.com   {az2407, sc250}@columbia.edu
github.com/alirezazareian/ovr-cnn

Abstract

Despite the remarkable accuracy of deep neural networks in object detection, they are costly to train and scale due to supervision requirements. Particularly, learning more object categories typically requires proportionally more bounding box annotations. Weakly supervised and zero-shot learning techniques have been explored to scale object detectors to more categories with less supervision, but they have not been as successful and widely adopted as supervised models. In this paper, we put forth a novel formulation of the object detection problem, namely open-vocabulary object detection, which is more general, more practical, and more effective than weakly supervised and zero-shot approaches. We propose a new method to train object detectors using bounding box annotations for a limited set of object categories, as well as image-caption pairs that cover a larger variety of objects at a significantly lower cost. We show that the proposed method can detect and localize objects for which no bounding box annotation is provided during training, at a significantly higher accuracy than zero-shot approaches. Meanwhile, objects with bounding box annotation can be detected almost as accurately as supervised methods, which is significantly better than weakly supervised baselines. Accordingly, we establish a new state of the art for scalable object detection.

1. Introduction

Object detection is one of the most prominent applications of artificial intelligence, and one of the most successful tasks for deep neural networks. However, despite the tremendous progress in deep object detection, such as Faster R-CNN [32] and its impressive accuracy, training such models requires expensive and time-consuming human supervision. Particularly, one needs to manually annotate at least thousands of bounding boxes for each object category of interest. Although such efforts have been already made and there are valuable datasets publicly available, such as Open Images [21] and MSCOCO [25], these datasets cover
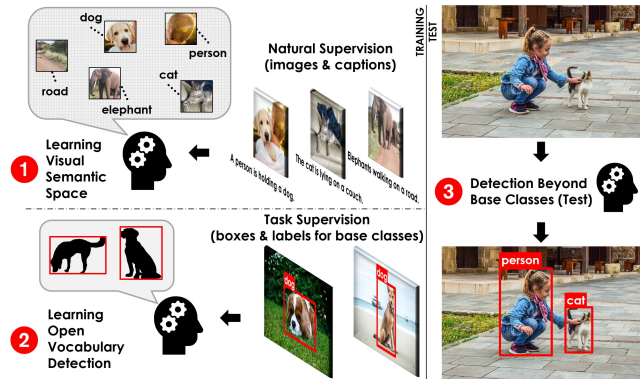

Figure 1. An overview of Open-Vocabulary Object Detection. We propose a two-stage training framework where we first (1) construct a visual-semantic space using low-cost image-caption pairs, and then (2) learn object detection using object annotations for a set of base classes. During test (3), the goal is to detect object categories beyond base classes, by exploiting the semantic space.

a limited set of object categories (*e.g.* 600), despite requiring extensive resources. Extending object detection from 600 to 60,000 categories requires 100 times more resources, which makes versatile object detection out of reach.

Nevertheless, humans learn to recognize and localize objects effortlessly through natural supervision, *i.e.*, exploring the visual world and listening to others describing situations. Their lifelong learning of visual patterns and associating them with spoken words results in a rich visual and semantic vocabulary that can be used not only for detecting objects, but for other tasks too, such as describing objects and reasoning about their attributes and affordances. Although drawing bounding boxes around objects is not a task that humans naturally learn, they can quickly learn it using few examples, and generalize it well to all types of objects, without needing examples for each object class.

In this paper, we imitate this human ability, by designing a two-stage framework named Open-Vocabulary object Detection (OVD). We propose to first use a corpus of image-caption pairs to acquire an unbounded vocabulary of

concepts, simulating how humans learn by natural supervision, and then use that knowledge to learn object detection (or any other downstream task) using annotation for only some object categories. This way, costly annotation is only needed for some categories, and the rest can be learned using captions, which are much easier to collect, and in many cases freely available on the web [33]. Figure 1 illustrates the proposed OVD framework, which is novel and efficient, enables versatile real-world applications, and can be generalized to other computer vision tasks.

More specifically, we train a model that takes an image and detects any object within a given *target* vocabulary $\mathcal{V}_T$. To train such a model, we use an image-caption dataset covering a large variety of words denoted as $\mathcal{V}_C$ as well as a much smaller dataset with localized object annotations from a set of *base* classes $\mathcal{V}_B$. Note that in this task, target classes <u>are not known</u> during training, and can be any subset of the entire language vocabulary $\mathcal{V}_\Omega$. This is in contrast with most existing object detection settings including weakly supervised transfer learning methods, where $\mathcal{V}_T$ should be known beforehand [37]. The most similar task to OVD is zero-shot object detection, which also generalizes to any given target set, but cannot utilize captions. Figure 2 illustrates an intuitive abstraction of our proposed task compared to zero-shot and weakly supervised detection. Despite close connections to those well-known ideas, OVD is novel and uniquely positioned in the literature, as we elaborate in Section 2.

To address the task of OVD, we propose a novel method based on Faster R-CNN [32], which is first pretrained on an image-caption dataset, and then fine-tuned on a bounding box dataset, in a particular way that maintains the rich vocabulary learned during pretraining, enabling generalization to object categories without annotation. Through extensive experiments, we evaluate our method, Open Vocabulary R-CNN (OVR-CNN), and show that it achieves significantly higher performance than the state of the art in zero-shot learning (27% mAP compared to 10%). We also show that it outperforms weakly supervised object detectors by a significant margin in generalized zero-shot settings (40% mAP compared to 26%). We supplement the paper with comprehensive open-source code to reproduce results.[1]

2. Related work

Zero-shot object detection (ZSD) aims to generalize from annotated (seen) object classes to other (unseen) categories. The key idea is to use zero-shot learning techniques (*e.g.* word embedding projection [13]) to learn object proposal classification. Bansal *et al.* [3] argued the main challenge in ZSD is modeling the background class, which is hard to separate from unseen classes. They defined background as a mixture model, which was later improved by
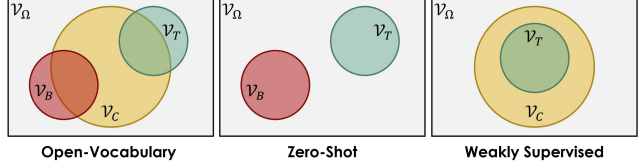
Figure 2. A comparison of our proposed OVD with existing ZSD and WSD paradigms. While zero-shot detection methods learn a limited set of base classes $\mathcal{V}_B$ and struggle to generalize to target classes $\mathcal{V}_T$, we acquire a much larger vocabulary $\mathcal{V}_C$ by learning from low-cost image-caption pairs. Although there are weakly supervised approaches that can learn from captions, they cannot use bounding box supervision from base classes, and need to know $\mathcal{V}_T$ before training. Hence, our OVD formulation is a generalization of ZSD and WSD, which can use both data sources to reach an outstanding performance on target classes not known in advance.

the introduction of polarity loss [30]. On the other hand, Zhu *et al.* [45, 46] argued the key to ZSD is to improve the generalization ability of object proposal models. They employed a generative model to hallucinate unseen classes and augment seen examples when training the proposal model. Nevertheless, ZSD methods are still far from practical performance, due to their unnecessarily harsh constraint, *i.e.*, not having any example of unseen objects, and having to guess how they look like solely based on their word embeddings [3, 30, 46] or textual descriptions [24]. This has motivated recent papers to simplify the task by making unrealistic assumptions, such as the availability of test data during training [29], or the availability of unseen class annotations to filter images with unseen object instances [15]. Considering datasets with natural, weak supervision are abundant and cheap, we propose an alternative, more realistic problem: Besides annotated data for "seen" classes, we assume an image-caption dataset is available that covers a larger variety of objects with an open vocabulary. This allows us to achieve 27% mAP on unseen classes, compared to the 10% state of the art, without much extra annotation effort. To this end, we address the open problem of knowledge transfer from image-caption pretraining to object detection.

Weakly supervised object detection (WSD) is the most widely used approach to train object detectors without bounding box annotations, by using image-level labels instead. The main challenge of WSD is localization, as each label may refer to any object in the image. This problem is typically addressed using multiple instance learning, which is a well-studied topic [4, 39, 8]. Although image-level labels are easier to collect than bounding boxes, they still require manual effort, and they are typically limited to a predefined taxonomy. In contrast, we use captions, which are more natural to annotate and often freely available on the web, while also featuring a rich, open vocabulary of concepts. Learning object detection from captions has been

studied at a limited scale. Cap2Det [43] parses captions into multi-label classification targets, which can be used to train a WSD model. However, that requires image-level labels to train the caption parser, and is limited to a closed vocabulary. Amrani *et al*. [2] train a WSD model based on the presence of a predefined set of words in captions, which is similarly closed-vocabulary, and discards the rich semantic content of captions, which we exploit through transformers. In contrast, Sun *et al*. [35] and Ye *et al*. [44] aim to discover an open set of object classes from image-caption corpora, and learn detectors for each discovered class. A key limitation of all such WSD methods is their inferior object localization accuracy. In contrast, we disentangle object recognition and localization into two independent problems. We learn recognition using open-vocabulary captions, while learning localization using a fully annotated dataset from a small subset of classes.

Object detection using mixed supervision has been studied in order to exploit both weak and full supervision. However, most existing methods need bounding box annotations for all classes, and use weak supervision only as auxiliary data [14, 31, 40]. More similar to our work are those which transfer a detector trained on supervised base classes to weakly supervised target classes [18, 36, 37]. These methods usually lose performance on base classes as we show in Section 4. In contrast, we treat this problem as an opposite knowledge transfer process: Instead of training on base classes first, and transferring to target classes using weakly supervised learning, we first use captions to learn an open-vocabulary semantic space that includes target classes, and transfer that to the task of object detection via supervised learning. Another limitation of all weakly supervised and mixed-supervision methods is that they require image-level annotations within a predefined taxonomy, and they only learn those predefined classes. In contrast, we use captions which are open-vocabulary and also more prevalent on the web, and we learn to generalize to any set of target classes on demand, without having to know them beforehand. VirTex [11] is the only method that uses captions as well as object annotations to train a detector, but it needs annotation for all object classes while we can generalize from a subset of annotated categories.

Visual grounding of referring expressions can be seen as an open-vocabulary object localization problem: Given an image and a noun phrase that refers to an object, usually within the context of a full caption, the goal is to localize the referred object in the image using a bounding box. We are inspired by the rich literature of weakly supervised visual grounding methods [42, 9, 5, 1] to design our image-caption pretraining technique. More specifically, we learn to map caption words to image regions, by learning a visual-semantic common space. However, such a mapping alone

cannot be used to detect objects during inference when no caption is provided. Therefore, we propose to transfer visual grounding knowledge to the task of object detection through another phase of training.

Vision-language transformers Our framework of pretraining with image-captions and transferring the learned knowledge to the downstream task is inspired by the recent success of multimodal transformers [26, 22, 23, 7, 34]. These methods train transformers in a self-supervised manner to take image-caption pairs as input and extract versatile features that can be fine-tuned on various downstream vision-language tasks. However, they have not been applied to object detection yet, since they need both image and caption as input, and also because they rely on a pretrained object detector to articulate the image before feeding into transformers. Recently, PixelBERT [19] removed the latter requirement by applying transformers directly on the feature map. We utilize and extend PixelBERT by devising a vision-to-language projection module before feeding visual features into the transformer, and by incorporating a visual grounding cost function into the pretraining process. Moreover, we propose to transfer the weights pretrained via multi-modality transformers to the single-modality downstream task of object detection.

3. Method

Figure 3 illustrates the architecture of our proposed method, which is based on a Faster R-CNN [32] trained in a zero-shot manner. More specifically, it is trained on a set of *base* classes $\mathcal{V}_B$, and tested on another set of *target* classes $\mathcal{V}_T$. To this end, pretrained word embeddings (*e.g.* GloVE [28]) are often used instead of conventional, trainable classifiers, so that target class embeddings can replace base class embeddings during testing, without changing the model's output semantic space. Nevertheless, this practice often leads to severe overfitting due to the small sample of base classes, to the point where the state-of-the-art mAP on target classes is 9 times lower than base classes [30].

To alleviate this problem, our key idea is to pretrain the visual backbone on a larger vocabulary $\mathcal{V}_C$ to learn a more complete semantic space rather than a small number of base classes. Since captions are naturally written without much constraint on the vocabulary, they are a perfect source for learning a rich and complete visual-semantic space. We name this framework Open Vocabulary Object Detection (OVD), as there are no explicit limits on the vocabulary of objects that can be learned through captions. In practice, our vocabulary is not literally "open", as it is limited to pretrained word embeddings. However, word embeddings are typically trained on very large text corpora such as Wikipedia that cover nearly every word [28, 12].

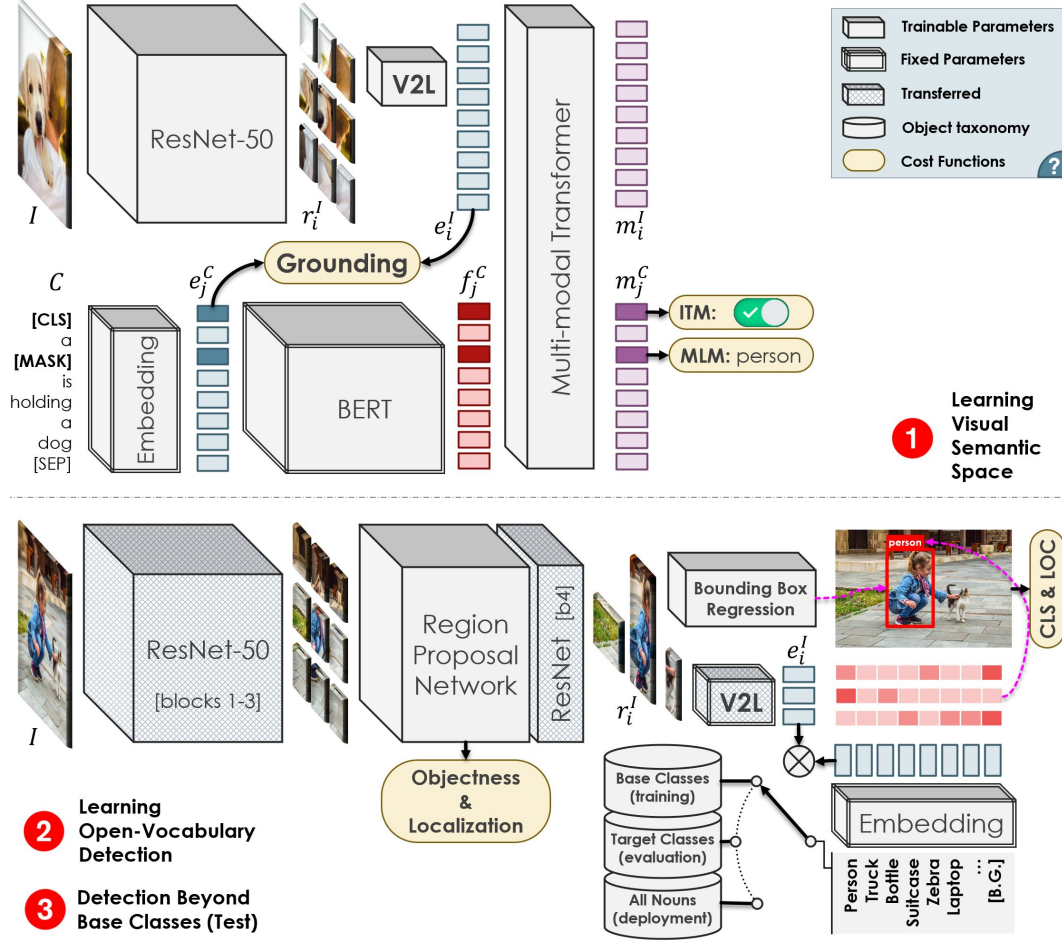In the rest of this section, we elaborate how we pretrain

Figure 3. The architecture of our OVR-CNN during pretraining (top) and downstream training (bottom). We first train the ResNet and the V2L layer on image-caption pairs via grounding, masked language modeling (MLM) and image-text matching (ITM). Then we use the trained ResNet and V2L to initialize a Faster R-CNN in order to learn open-vocabulary object detection.

our Open Vocabulary faster R-CNN (OVR-CNN) on image-caption pairs, and how we transfer the pretraining knowledge to the downstream task. In Section 4, we demonstrate that our method closes the base-target performance gap from a ratio of 9 to 2.

3.1. Learning a visual-semantic space

Object detectors typically use a CNN backbone that is often pretrained for ImageNet classification [10, 32]. Pretraining results in a backbone that can extract features optimized for object recognition, which is then used to train a new classification head for a fixed set of annotated classes. This is problematic in zero-shot settings, as a classifier trained on base classes cannot recognize target classes. Therefore, zero-shot methods learn a linear projection from visual features to pretrained base class embeddings by replacing classifier weights with a fixed embeddings matrix [13]. This way, the network is expected to generalize to

target classes by assuming the continuity of the embedding space. Nevertheless, this approach is prone to overfitting, as projecting to a small number of the embedding space (base class embeddings) is an under-determined problem [3].

To prevent overfitting, we propose to learn the aforementioned Vision to Language (V2L) projection layer along with the CNN backbone during pretraining, where the data is not limited to a small set of base classes. To this end, we use an image-caption dataset, since captions contain a rich vocabulary and semantic structure that can be used to learn the meaning of words, including object names. To effectively learn from the rich supervision that captions provide, we exploit recent advances in visual grounding and vision-language transformers. We use a main (grounding) task as well as a set of auxiliary self-supervision tasks to learn a robust CNN backbone and V2L layer. In the next subsection, we elaborate how we transfer the pretrained modules to learn open-vocabulary object detection.

Our pretraining architecture resembles PixelBERT [19]: it takes image-caption pairs as input, feeds the image into a visual backbone and the caption into a language backbone, which results in a set of token embeddings for the image and caption, and then feeds those token embeddings into a multimodal transformer to extract multimodal embeddings. Our visual backbone is a ResNet-50 [17], which takes a $w \times h$ image I as input and extracts a grid of $w/32 \times h/32$ regions, where each region i is represented by a d_v-dimensional feature vector, r_i^I. For the language backbone, we use a pre-trained BERT [12], which takes a tokenized caption C as input, extracts a d_l-dimensional word embedding e_j^C for each token j, augments that with position embeddings, and applies several layers of multi-head self-attention to extract d_l-dimensional contextualized token embeddings f_j^C.

Furthermore, we devise a linear V2L layer that maps each visual region representation r_i^I into the language embedding space e_i^I. The final embeddings of image regions $\{e_i^I\}$ and caption tokens $\{f_j^C\}$ are then aggregated and fed into a multimodal transformer, which is similar to BERT in architecture, but performs attention not only within each modality but also across the two modelities. The output of the multimodal transformer is $\{m_i^I\}$ and $\{m_j^C\}$ for the regions and words respectively, which can be used for various pretraining tasks, as we discuss later in this section.

Once we extract the aforementioned stages of unimodal and multimodal embeddings from a batch of image-caption pairs, we define a main objective function as well as various auxiliary objectives to ensure an effective training for the ResNet parameters, as well as the V2L layer. Our main objective is visual grounding, *i.e.*, word embeddings from each caption e_j^C should be close to their corresponding image regions e_i^I. Since the correspondence of words and regions is not given, we employ a weakly supervised grounding technique to learn it. Specifically, we define a global grounding score for each image-caption pair, that is a weighted average of local grounding scores for word-region pairs:

$$\langle I, C \rangle_G = \frac{1}{n_C} \sum_{j=1}^{n_C} \sum_{i=1}^{n_I} a_{i,j} \langle e_i^I, e_j^C \rangle_L, \qquad (1)$$

where $\langle ., . \rangle_L$ is the dot product of two vectors, n_I and n_C are the number of image and caption tokens, and

$$a_{i,j} = \frac{\exp\langle e_i^I, e_j^C \rangle_L}{\sum_{i'=1}^{n_I} \exp\langle e_{i'}^I, e_j^C \rangle_L}. \qquad (2)$$

The global grounding score for a matching image-caption pair should be maximized, while it should be minimized for a non-matching pair. Hence, we use other images in the batch as negative examples for each caption, and use other captions in the batch as negative examples for each image.

Accordingly, we define two grounding objective functions:

$$\mathcal{L}_G(I) = -\log \frac{\exp\langle I, C \rangle_G}{\sum_{C' \in \mathcal{B}_C} \exp\langle I, C' \rangle_G}, \qquad (3)$$

and

$$\mathcal{L}_G(C) = -\log \frac{\exp\langle I, C \rangle_G}{\sum_{I' \in \mathcal{B}_I} \exp\langle I', C \rangle_G}, \qquad (4)$$

where $\mathcal{B}_I$ and $\mathcal{B}_C$ are the image and caption batch. We validated the described formulation by completing extensive experimentation with various other alternatives, such as other similarity metrics (*e.g.* cosine instead of dot product), other loss functions (*e.g.* triplet loss instead of negative log likelihood) and other word-to-region alignment mechanisms (*e.g.* hard alignment instead of softmax).

Optimizing the grounding objectives results in a learned visual backbone and V2L layer that can map regions in the image into words that best describe them, without limiting to a closed vocabulary. However, since we induce a weak, indirect supervision, a local optima might be achieved where the model only learns the minimum concepts necessary to choose the right image/caption. To more directly learn each word, we employ masked language modeling following PixelBERT [19]. Specifically, we randomly replace some words j in each caption C with a [MASK] token, and try to use the multimodal embedding of the masked token m_j^C to guess the word that was masked. To this end, the visual backbone and the V2L layer should learn to extract all objects that might be described in captions, and the multimodal transformer should learn to use those along with the language understanding ability of BERT to determine what word completes the caption best.

Accordingly, we apply a fully connected layer on m_j^C, compare its output to all word embeddings using dot product, and apply softmax to compute a probability score for each word. We define masked language modeling $\mathcal{L}_{MLM}$ as a cross-entropy loss comparing the predicted distribution with the actual word that was masked. PixelBERT also employs an image-text matching loss $\mathcal{L}_{ITM}$, but does not use masked visual modeling that is common in vision-language transformers [26]. We follow their choices for our auxiliary objectives, although other combinations are possible. We train the visual backbone, V2L layer, and the multimedia transformer jointly by minimizing the total loss for each image-caption pair:

$$\mathcal{L}(I, C) = \mathcal{L}_G(I) + \mathcal{L}_G(C) + \mathcal{L}_{MLM} + \mathcal{L}_{ITM}. \qquad (5)$$

Note that our language backbone (BERT) and its word embeddings are fixed in our experiments.

3.2. Learning open-vocabulary detection

Once the ResNet visual backbone and V2L layer are trained, we transfer them to the task of object detection,

by initializing a Faster R-CNN. Following [32], we use the stem and the first 3 blocks of our pretrained ResNet to extract a feature map from a given image. Next, a region proposal network slides anchor boxes on the feature map to predict objectness scores and bounding box coordinates, followed by non-max suppression and region-of-interest pooling to get a feature map for each potential object. Finally, following [32], the 4th block of our pretrained ResNet is applied on each proposal followed by pooling to get a final feature vector r_i^I for each proposal box, which is typically fed into a classifier in supervised settings.

Nevertheless, in our zero-shot setting, a linear layer is applied on the visual features r_i^I to map each proposal onto a word embedding space e_i^I, so they can be compared to base or target class embeddings in the training or testing phase respectively. In all ZSD methods, the aforementioned linear layer is trained from scratch on base classes, which struggles to generalize. In contrast, we have already trained the V2L layer in the pretraining phase, on a much broader semantic space. The main difference of this phase with pretraining is that instead of the grid-structured feature map, r_i^I represents a bounding box of arbitrary shape. However, due to the linear characteristics of RoI-Align [16], r_i^I is on the same space as in pretraining, with minimal domain shift that can be eliminated by fine-tuning the ResNet backbone.

During training, we compare e_i^I to each base class k to compute classification scores:

$$p(i \text{ classified as } k) = \frac{\exp\langle e_i^I, e_k^\mathcal{V}\rangle}{1 + \sum_{k'\in\mathcal{V}_B} \exp\langle e_i^I, e_{k'}^\mathcal{V}\rangle}, \quad (6)$$

where $e_k^\mathcal{V}$ is the pretrained embedding of word k, $\mathcal{V}_B$ is the set of base classes, and $\langle .,.\rangle$ denotes dot product. The addition of 1 in the denominator is because we set the background class to have a fixed, all-zero embedding, which makes any dot product zero, and is exponentiated to 1. We found that a fixed all-zero background embedding performs better than a trainable one as it does not push non-foreground bounding boxes, which may contain target classes, to an arbitrary region of the embedding space.

Except for the aforementioned changes in the classification head, the rest of our network exactly follows Faster R-CNN, and is trained in the exact same way with the same objective functions. Empirically, we found that multiplying a ratio α to the classification loss of background proposals (*i.e.*, proposal boxes that are not matched with any ground truth bounding box) can improve the performance on target classes significantly, while slightly lowering base class performance. Hence, we use cross-validation to find the best α for each model. The ResNet parameters are finetuned, while the region proposal network and the regression head are trained from scratch. The classifier head is fully fixed, as it consists of a pretrained V2L layer and word embeddings, which are especially prone to overfitting. During testing,

we use the model just like a Faster R-CNN, except we can replace word embeddings in Eq. (6) with any set of target classes $\mathcal{V}_T$. While we evaluate on a fixed, annotated target set, the model is not particularly tuned for those classes, and hence can be deployed on the entire vocabulary $\mathcal{V}_\Omega$.

4. Experiments

In this section, we demonstrate our method's ability to detect objects of the target classes accurately, while not losing its accuracy on the base classes compared to supervised approaches. Particularly, we show significant quantitative improvements compared to zero-shot and weakly supervised object detection methods, followed by a comprehensive analysis including ablation and visualization.

4.1. Data and metrics

We base our experiments on the challenging and widely used COCO Objects dataset [25]. We use their 2017 training and validation split for training and evaluation respectively. To select base and target classes, we adopt the split proposed by [3] and used by all other ZSD methods. Their splits includes 48 base classes and 17 target classes, which are both subsets of COCO object classes. We remove any bounding box that is not labeled with a base class from training data, and remove images that are left with no bounding boxes. This leaves us with 107,761 training images that contain 665,387 instances of base classes, and 4,836 test images that contain 28,538 instances of base classes and 4,614 instances of target classes.

Unless otherwise mentioned, for pretraining we use COCO Captions [6], which is based on the same images and same train/test split as COCO Objects. This dataset is preferred due to the matching domain with the downstream task. However, to study more realistic settings, we also report results by pretraining on Conceptual Captions (CC) [33], which was automatically collected from the web. CC is larger with 2,749,293 training image-caption pairs, compared to COCO with 118,287 images and 5x captions. Both COCO and CC cover very broad vocabularies that include all base and target classes in our experiments. Although there is no theoretical limit for our model to predict words outside the caption vocabulary $\mathcal{V}_C$ (Figure 2), we do not study those unusual cases in this paper.

Following most ZSD and WSD methods, we evaluate using mean Average Precision (mAP) at an IoU of 0.5. We compute mAP on base classes by directly applying the model on COCO validation images and using base class annotations to evaluate. Then we replace the classifier head with target class embeddings and apply on the same dataset, but this time compare with target class annotations. These result in base and target mAP, which resemble supervised and zero-shot settings respectively. We also replace the classifier head with the union of base and target class em-

beddings, to mimic generalized zero-shot settings [30]. In that case, we report total mAP, as well as separately computing the mean of AP over base and target classes.

4.2. Implementation details

We used the `maskrcnn-benchmark` code base [27], and particularly the `R_50_C4` configuration to implement our system. We also used a pretrained and frozen `BERT-Base` [41] as our language backbone. For the multimodal transformer, we use the same architecture as `BERT-Base`, except we use only 6 layers and 8 attention heads at each layer, and we train it from scratch. Our base learning rate for pretraining is 0.01 which drops to 0.001 and 0.0001 after sufficient training. We use a batch size of 64 and train on 8 V-100 GPUs which takes about 10 hours. We use spatial dropout following [19] to subsample visual regions during pretraining. For masked language modeling, we mask each word with the likelihood of 0.135. We use gradient clipping at 5.0 for pretraining.

During downstream training, we use the BERT embeddings (*i.e.*, pretrained input embeddings, not the output of BERT transformers) of the base classes to initialize and fix the classifier weights. We found the best background weight is $\alpha = 0.2$ for most experiments, except the ablations without a fixed, pretrained V2L layer, where $\alpha = 0.0$ works best. We only fine-tune the third and forth block of ResNet, and keep the stem and first two blocks fixed. We train using a learning rate of 0.005 and drop to 0.0005 and 0.00005 when appropriate. We train with a batch size of 8 on 8 V-100 GPUs which takes about 18 hours to converge.

4.3. Baselines

Because our proposed OVD task utilizes a unique combination of training data and supervision that has not been studied before, there are no baselines with identical training settings for an entirely fair comparison. Therefore, we use baselines from a variety of similar but not identical tasks. Firstly, we compare to zero-shot detection methods, as ZSD is the closest area to our work. Particularly, we compare to SB [3], which is the first and simplest ZSD method, projecting CNN features of EdgeBox proposals [47] to word embeddings. Then we compare to LAB [3], which attempts to better model the background class using a mixture model. We also compare to DSES [3], which uses additional classes from Visual Genome [20] to augment base classes. Then we compare to PL [30], which proposes polarity loss to address the object-background imbalance, and to DELO [46], which employs a generative approach to prepare the model for certain target classes through feature hallucination. Note that DELO needs to know target classes beforehand, which makes it not truly open-vocabulary.

It is important to note that our approach utilizes extra data (COCO Captions or Conceptual Captions) that is

Method	Task	Base (48)	Target (17)	Generalized (48+17)		
				Base	Target	All
FR-CNN [32]	FSD	54.5	-	-	-	-
WSDDN [4]*	WSD	-	-	19.6	19.7	19.6
Cap2Det [43]*		-	-	20.1	20.3	20.1
LSDA [18]*	MSD	-	-	29.3	17.7	27.2
LSDA+[36]*		-	-	28.5	21.9	26.7
MIL+RPN[37]*		-	-	27.8	22.6	26.4
SB [3]	ZSD	29.7	0.70	29.2	0.31	24.9
LAB [3]		21.1	0.27	20.8	0.22	18.0
DSES [3]		27.2	0.54	26.7	0.27	22.1
DELO [46]*		14.0	7.60	13.8	3.41	13.0
PL [30]		36.8	10.0	35.9	4.12	27.9
OVR-CNN	**OVD**	**46.8**	**27.5**	**46.0**	**22.8**	**39.9**

Table 1. Results on the MSCOCO dataset. Numbers are mAP (%). *For some baselines, target classes are known during training.

not available to ZSD baselines, and may include examples of target categories. Therefore, we also compare to weakly supervised detection (WSD) methods, by converting captions into image-level labels using exact matching or a classifier [43]. We compare to WSDDN [4], as well as Cap2Det [43] which better utilizes captions. WSD methods do not utilize bounding boxes for base classes, which can be an advantage in situations where no such annotation is available, but it results in poor localization performance compared to our method, which is able to utilize bounding boxes. Hence, we also compare to transfer learning methods that utilize a mixture of weak and full supervision (denoted as MSD). Particularly, we compare to LSDA [18], which learns a transformation from classifier weights into detector weights, its extension [36] to utilize semantic class relationships (LSDA+), and a more recent work [37] which uses multiple instance learning along with a region proposal network trained on base classes (MIL+RPN).

Note that since WSD and MSD methods require image-level labels, target classes should be known in advance during pretraining, and the models are particularly adapted to those classes. In contrast, our method and most ZSD methods have no access to such information, and can be applied to any novel class without retraining.

4.4. Results

Table 1 demonstrates our main results compared to the baselines. Particularly, we observe a significant improvement on target class performance and generalized target performance compared to all ZSD baselines. This is mainly due to our ability to utilize additional, low-cost training data. We also outperform WSD and MSD baselines on target classes, despite their access to information about target classes during training, and we significantly outperform them on base classes and therefore overall, due to our effective exploitation of bounding box supervision for base classes. Note that WSD and MSD models cannot be evaluated on base-only or target-only classes since they have a

Ablation	Base (48)	Target (17)	All (65)
Ours w/o pretraining	25.2	4.4	18.1
Ours w/o grounding	25.9	4.6	19.0
Ours w/o auxiliary objectives	45.6	26.0	38.8
Ours w/o transferring V2L	25.3	4.9	18.6
Ours w/o freezing V2L	47.0	23.4	39.3
Ours w/o Imagenet	18.4	9.13	14.3
Ours w/ Conceptual Captions	43.0	16.7	34.3
Ours	**46.8**	**27.5**	**39.9**

Table 2. Ablation on MSCOCO dataset. Numbers are mAP (%).

fixed classifier trained on all 65 classes. Moreover, we have a FSD (fully supervised detection) baseline to measure the performance drop on base classes.

Furthermore, we present ablation experiments in Table 2 to show the effectiveness of each design choice. Particularly, we observe that without pretraining our model on image-caption datasets, the model performs poorly. This confirms the remarkable efficacy of multimodal pretraining for open-vocabulary generalization. We also observe that grounding is the main component of pretraining, which has a much larger effect than the auxiliary objectives that are optimized through the multimedia transformer module. Moreover, we show that transferring ResNet weights alone (from pretraining to downstream task) is not enough for effective knowledge transfer, and we must transfer the V2L layer as well. Additionally, if the V2L layer is not frozen during downstream training, it loses its ability to generalize to target classes, in order to slightly improve on base classes.

We also try initializing the model randomly during pretraining instead of using widely used Imagenet weights, and despite the performance drop, we still perform better than most ZSD baselines that use Imagenet. We also observe that if we use the automatically collected Conceptual Captions instead of the carefully annotated COCO Captions, the performance drops, but still outperforms all ZSD baselines significantly, proving that even low-quality, cheap data can be utilized by OVR-CNN to achieve better performance.

4.5. Visualization and discussion

To gain deeper insight about what OVR-CNN learns, we depict the visual-semantic embedding space that is learned by our model in Figure 4. More specifically, we apply our trained model (after downstream training) on all COCO validation images, get the embeddings of all output bounding boxes after the V2L layer e_i^I, and reduce their dimensionality to 2 using t-SNE [38]. We color-code them based on their ground truth label and overlay class embeddings e_k^V on the same space. We only show target classes and their instances to reduce clutter. Ideally, instances of each target class should form distinct clusters, and each class embed-
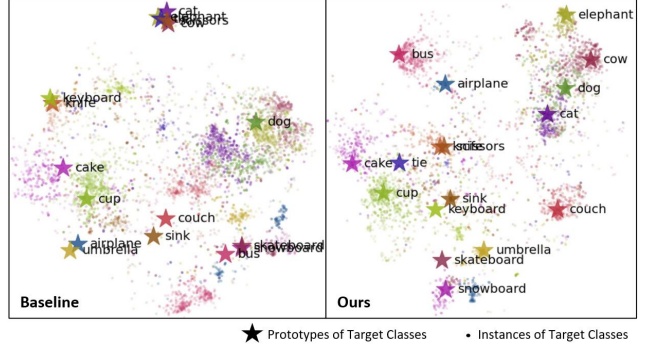

Figure 4. The embedding space learned by OVR-CNN (right) compared to a baseline without pretraining (left). Each color represents a target class, each dot represents the e_i^I embedding of a bounding box and each star represents a class prototype.

ding (prototype) should fall inside the cluster formed by its instances. This is particularly difficult to achieve for target classes due to the lack of direct supervision. We compare our method to a ZSD baseline that is identical to our model except without pretraining on image-caption pairs.

We observe that in the baseline, target classes form convoluted clusters and their prototypes are randomly distributed or collapsed. On the other hand, our full model creates well-defined clusters that contain their prototypes in most cases. This is consistent with our intuition and our quantitative results that suggest zero-shot learning is not sufficient for learning a smooth and generalizable mapping from visual features to semantic embeddings, and learning a larger vocabulary through multimodal data is crucial for a more coherent space and generalizing beyond base classes.

5. Conclusion

We called attention to the new task of Open-Vocabulary Object Detection (OVD), as an attempt to disentangle object detection into recognition and localization, and learn them separately using two different sources of supervision that are perfect for each corresponding task. In OVD, recognition is learned from captions, which are general-purpose and open-vocabulary, while localization is learned from bounding box annotations, which are accurate and directly designed for the downstream task. We proposed OVR-CNN which pretrains a Faster R-CNN on an image-caption dataset and carefully transfers the open-vocabulary visual-semantic knowledge learned from captions to the downstream task of object detection. We demonstrated record performance compared to zero-shot and weakly supervised baselines, establishing a new state of the art for scalable object detection. Nevertheless, OVR-CNN is merely one possible implementation of our general idea, which can be extended to other downstream tasks too, enabling more human-like, open-vocabulary computer vision technology.

References

[1] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12476–12486, 2019. 3

[2] Elad Amrani, Rami Ben-Ari, Tal Hakim, and Alex Bronstein. Learning to detect and retrieve objects from unlabeled videos. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3713–3717. IEEE, 2019. 3

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 2, 4, 6, 7

[4] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 2, 7

[5] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050, 2018. 3

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[7] 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 *ECCV*, 2020. 3

[8] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multifold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2016. 2

[9] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2601–2610, 2019. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[11] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 5

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 2, 4

[14] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. Note-rcnn: Noise tolerant ensemble rcnn for semisupervised object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9508–9517, 2019. 3

[15] Dikshant Gupta, Aditya Anantharaman, Nehal Mamgain, Vineeth N Balasubramanian, CV Jawahar, et al. A multi-space approach to zero-shot object detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1209–1217, 2020. 2

[16] 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. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014. 3, 7

[19] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 3, 5, 7

[20] 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*, 123(1):32–73, 2017. 7

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, pages 1–26, 2020. 1

[22] 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. 3

[23] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Weaklysupervised visualbert: Pre-training without parallel images and captions. *arXiv preprint arXiv:2010.12831*, 2020. 3

[24] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8690–8697, 2019. 2

[25] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 6

[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. 3, 5

[27] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: [Insert date here]. 7

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3

[29] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6082–6091, 2019. 2

[30] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. *34th AAAI Conference on Artificial Intelligence*, 2020. 2, 3, 7

[31] Vignesh Ramanathan, Rui Wang, and Dhruv Mahajan. Dlwl: Improving detection for lowshot classes with weakly labelled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9342–9352, 2020. 3

[32] 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*, pages 91–99, 2015. 1, 2, 3, 4, 6, 7

[33] 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. 2, 6

[34] 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 *International Conference on Learning Representations*, 2020. 3

[35] Chen Sun, Chuang Gan, and Ram Nevatia. Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE international conference on computer vision*, pages 2596–2604, 2015. 3

[36] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2016. 3, 7

[37] Jasper Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2018. 2, 3, 7

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 8

[39] 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 Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019. 2

[40] Yu-Xiong Wang and Martial Hebert. Model recommendation: Generating object detectors from few samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1619–1628, 2015. 3

[41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019. 7

[42] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954, 2017. 3

[43] 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 International Conference on Computer Vision*, pages 9686–9695, 2019. 3, 7

[44] Keren Ye, Mingda Zhang, Wei Li, Danfeng Qin, Adriana Kovashka, and Jesse Berent. Learning to discover and localize visual objects with open vocabulary. *arXiv preprint arXiv:1811.10080*, 2018. 3

[45] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Zero shot detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):998–1010, 2019. 2

[46] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702, 2020. 2, 7

[47] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014. 7