Learning to Overcome Noise in Weak Caption Supervision for Object Detection

Mesut Erhan Unal, Keren Ye, Mingda Zhang, Christopher Thomas, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent

Abstract—We propose the first mechanism to train object detection models from weak supervision in the form of captions at the image level. Language-based supervision for detection is appealing and inexpensive: many blogs with images and descriptive text written by human users exist. However, there is significant noise in this supervision: captions do not mention all objects that are shown, and may mention extraneous concepts. We first propose a technique to determine which image-caption pairs provide suitable signal for supervision. We further propose several complementary mechanisms to extract image-level pseudo labels for training from the caption. Finally, we train an iterative weakly-supervised object detection model from these image-level pseudo labels. We use captions from four datasets (COCO, Flickr30K, MIRFlickr1M, and Conceptual Captions) whose level of noise varies. We evaluate our approach on two object detection datasets. Weighting the labels extracted from different captions provides a boost over treating all captions equally. Further, our primary proposed technique for inferring pseudo labels for training at the image level, outperforms alternative techniques under a wide variety of settings. Both techniques generalize to datasets beyond the one they were trained on.

Index Terms—Language-supervised object detection, weakly-supervised object detection, vision and language

1 Introduction

Learning to localize and classify objects in images is a fundamental problem in computer vision. It has a wide range of applications, including robotics, autonomous vehicles, intelligent video surveillance, and augmented reality. Modern detectors are highly accurate [1], can run in real-time [2] and on mobile devices [3]. Despite these achievements, most modern detectors suffer from an important limitation: they are trained with expensive supervision in the form of large quantities of bounding boxes meticulously drawn by a large pool of human annotators. Due to the well-known domain shift problem [4], [5] and imperfect domain adaptation techniques, this means that when detection is to be performed in a novel domain, the expensive annotation procedure needs to be repeated.

Weakly supervised object detection (WSOD) techniques aim to alleviate the burden of collecting expensive box annotations. The classic WSOD formulation [6], [7], [8] treats an image as a bag of proposals, and learns to assign instance-level semantics to these proposals. WSOD has shown great potential for object detection, and recent methods have reached 52% mAP [9] on Pascal VOC 2012.

However, we highlight two limitations of WSOD methods. First, they depend on large-scale image-level object category labels; these require human effort that is provided in an unnatural, crowdsourced environment. Second, they make the assumption that the image-level label should be *precise*, i.e., at least one proposal instance in the image needs to be associated with the label. This assumption does not hold for real-world problems and real-world supervision.

We propose mechanisms to leverage a new form of supervision for training weakly-supervised object detectors, namely supervision in the form of natural language descriptions that web users provide when uploading their photos to social media sites such as Instagram, or their videos to video sharing platforms such as YouTube. There are tens of millions of photos uploaded to Instagram every day, and a majority have titles, tags, or descriptions. Abundant videos with subtitles or descriptive narratives are similarly available on YouTube. These annotations are "free" in that no user was *paid* to provide them; they arise out of innate needs of users to make their content available to others.[1]

However, existing WSOD methods cannot use such supervision. First, *natural language descriptions are unstructured*; they need to be parsed and words relevant for object recognition to be extracted, while non-object words are removed. Second, these descriptions are *both imprecise and non-exhaustive*—they might mention content that is not in the image (e.g., what event the user was attending or who

- *Mesut Erhan Unal, Keren Ye, Mingda Zhang, and Adriana Kovashka are with the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260 USA. E-mail: meu6@pitt.edu, {yekeren, mzhang, kovashka}@cs.pitt.edu.*
- *Christopher Thomas is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA. E-mail: chris@cs.pitt.edu.*
- *Wei Li is with NewsBreak, Seattle, WA 98004 USA. E-mail: lwthucs@gmail.com.*
- *Danfeng Qin and Jesse Berent are with the Google Research, 8002 Zurich, Switzerland. E-mail: {qind, jberent}@google.com.*

Manuscript received 22 February 2021; revised 25 March 2022; accepted 19 June 2022. Date of publication 30 June 2022; date of current version 6 March 2023.
This work was supported in part by the National Science Foundation under Grants 1566270, 1718262, and 2046853, and in part by Google Faculty Research Awards and an NVIDIA hardware grant.
(Corresponding author: Adriana Kovashka.)
Recommended for acceptance by J. Hays.
Digital Object Identifier no. 10.1109/TPAMI.2022.3187350

[1] Of course, this data may be subject to license agreements limiting uses, and not all of it can truly be used for "free."

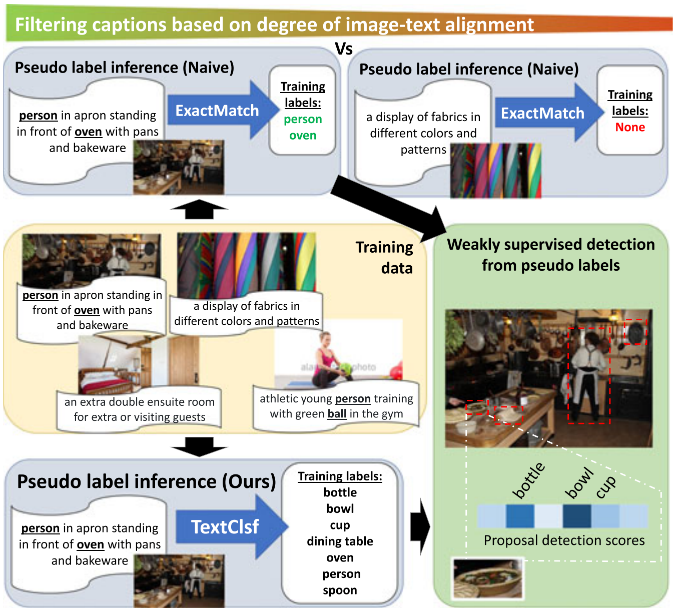

Filtering captions based on degree of image-text alignment
Vs
Pseudo label inference (Naive)
person in apron standing in front of oven with pans and bakeware
ExactMatch
Training labels:
person
oven
Pseudo label inference (Naive)
a display of fabrics in different colors and patterns
ExactMatch
Training labels:
None
Training data
person in apron standing in front of oven with pans and bakeware
a display of fabrics in different colors and patterns
an extra double ensuite room for extra or visiting guests
athletic young person training with green ball in the gym
Weakly supervised detection from pseudo labels
bottle
bowl
cup
Proposal detection scores
Pseudo label inference (Ours)
person in apron standing in front of oven with pans and bakeware
TextClsf
Training labels:
bottle
bowl
cup
dining table
oven
person
spoon

Fig. 1. We propose two mechanisms to infer pseudo training labels from captions. First (top), we determine the potential for strong object supervision signal from image-caption pairs (showing one with strong and one with weak signal). When supervision is strong, a simple training label extraction technique can be used. Second (bottom), we learn a mapping function (a text classifier) from captions to labels, which compensates for failures of exact-matching label extraction. Finally (right), we train a weakly-supervised object detection model with these pseudo image labels.

they met after the photo was taken), and omit content that is in the image but is not interesting. In the bottom of Fig. 1, many large objects—e.g., dining table and bowls—were not mentioned in the human-provided description. Thus, directly feeding web data to the state-of-the-art WSOD system is infeasible, which under-utilizes the rich supervision that language on the web can provide.

To address this issue, we propose a three-part framework to build an object detector from images paired with accompanying captions (sentences). Our model bridges human-written free-form texts and visual objects, and generates accurate bounding boxes over objects in an image. Our key contributions are the first two steps, with a smaller contribution in the third step. First, we *estimate which image-caption pairs can serve as appropriate supervision* for extracting pseudo *image-level* training labels for training an object detector. In particular, we model the difference between images which are visual neighbors and those which appear with similar captions (semantic neighbors). We prioritize extracting signal from image-caption pairs where visual and semantic neighborhoods overlap, which indicates that captions closely follow the image. This enables the use of simple techniques for extracting training labels (Fig. 1 top).

Second, we devise complementary *advanced* techniques for *extracting pseudo image-level training labels from the caption*. One of our proposed strategies (Fig. 1 bottom) is to train a textual classifier to map captions to discrete object labels. Unlike the previous contribution, this classifier requires a small set of labels, and enables us to bridge the gap between what humans mention in a caption, and what truly is in an image. Alternatives include learning multimodal spaces where images and captions are projected, and using similarity in these spaces to determine which captions are similar to object words in a predefined vocabulary. This contribution and the previous have different applications: The former is fitting when no labels are available, but if they are, the latter achieves slightly stronger performance. Thus, we primarily focus on evaluating these contributions separately, as shown in Fig. 1. Both contributions generalize beyond dataset boundaries.

Third, we use the *pseudo ground truth labels at the image level* (extracted in the previous step), *to train a weakly supervised object detection method*. The method we propose extracts region proposals off-the-shelf, then for each proposal and each class, learns both a class score and a detection score. These scores are then refined using an iterative approach, and combined to produce final detection results.

In our work, we first need to infer image-level pseudo labels for training from the available captions. Only then can we proceed to train a weakly-supervised detection (WSOD) algorithm, using those (potentially noisy) image-level pseudo labels. Thus, to distinguish our work from WSOD, we refer to our methods as performing *language-supervised object detection (LSOD)*.

Our paper makes four main contributions. First, we propose a new task of learning from noisy caption annotations, and set up a new benchmark. Rather than treating object categories as IDs only, we also leverage their semantics and synonyms of those object names. Second, we show the impact of multiple possible ways to map captions to image-level labels, ranging from exactly matching the captions to object category words, using learned image-text similarity scores, retrieving hand-annotated or predicted synonyms to the object categories from the captions, or training a classifier. Our proposed approach outperforms the baseline by up to 78% on noisy datasets. Third, we demonstrate the success of explicitly modeling which image-caption pairs provide strong signal for supervision, using a new metric that captures how closely the text follows the image. This alignment metric allows us to improve performance by up to 37%. Fourth, we show cross-domain results in datasets: we not only demonstrate competitive WSOD performance by training/testing on COCO captions, but also validate the benefit of our COCO-trained text classifier and alignment metric by applying it on Flickr30K, and the noisy MIRFlickr1M and Conceptual Captions. We are not aware of other work that directly extracts labels for detection training from the latter two datasets (and only a few works pretrain for detection on Conceptual Captions). We leverage the resulting models and evaluate them on the PASCAL and COCO object detection datasets.

The remainder of the paper is organized as follows. We overview related work in Section 2. In Section 3, we discuss how to filter or weight image-caption pairs as potential signals for supervision (Section 3.1), different ways to reduce the gap between free-form captions and object categories (Section 3.2), and the backbone of our WSOD model, which combines prior work [8], [10] in a new way (Section 3.3). In Section 4, we compare to upper and lower bounds, in conjunction with state-of-the-art methods. We conclude in Section 5.

2 Related Work

We formulate a new variant of weakly-supervised object detection, which we term *language*-supervised, where the supervision is even more weak but less costly than in prior work. We leverage vision-language interactions, so we also discuss work that finds alignment between image regions and text and grounds language in images. We also discuss recent work in learning visual representations from language. Finally, we describe work that investigates what kind of content humans describe in captions or models how visually concrete particular words are. We are not aware of other work that *explicitly handles noise for language-supervised object detection*, as we propose.

2.1 Weakly-Supervised Object Detection

Weakly-supervised object detection (WSOD) involves localizing and categorizing objects without instance-level (bounding box) supervision. Key approaches include multiple-instance learning (MIL) where one or more regions are associated with the label of interest [8], [11], and self-training, where high-scoring proposals are treated as pseudo ground-truth [8], [12], [13]. In the multiple-instance learning (MIL) setting, proposals of an image are treated as a bag of candidate instances. If the image is labeled as containing an object, at least one of the proposals will be responsible to provide the prediction of that object. Oquab *et al.* [14] and Zhou *et al.* [15] propose a Global Average (Max) Pooling layer to learn class activation maps. Bilen *et al.* [6] propose Weakly Supervised Deep Detection Networks (WSDDN) containing classification and detection data streams, where the detection stream weighs the results of the classification predictions. Tang *et al.* [8], [16] jointly train multiple refining models together with WSDDN, and show the final model benefits from the online iterative refinement. Diba *et al.* [17] and Wei *et al.* [7] apply a segmentation map; Wei *et al.* [7] further incorporate saliency. Wan *et al.* [18] add a min-entropy loss to reduce the randomness of the detection results. Zeng *et al.* [12] jointly consider bottom-up and top-down objectness from low-level measurement and CNN confidences. Ren *et al.* [9] aim at instance-aware self-training where they design DropBlock to zero out the most discriminative parts to avoid the part domination issue in WSOD. Earlier (pre-deep-learning) approaches include Divvala *et al.* [19] which rely on web search for an initial set of concepts for which to learn detection models, prune them based on model performance, and combine synonyms.

Our work is similar to these since we also represent the proposals using a MIL-weighted representation. However, prior WSOD methods require structure in the form of class labels, and these labels require dedicated human effort. Our contribution is enabling weakly-supervised detection with less costly *language supervision* which could work without explicit human annotations. In this project, we use both crowdsourced captions (from the COCO and Flickr30K datasets) and noisier ones obtained as a side product of users uploading content on the web (MIRFlickr1M, Conceptual Captions). We explicitly handle the noise in the language supervision and the misalignment between nouns (objects) that are shown but not mentioned, or mentioned but not shown. This distinguishes our work from both WSOD and self-supervised methods.

2.2 Vision-Language Tasks

Learning visual-semantic embeddings (VSE) has received tremendous interest due to its broad applications such as retrieval [20], [21], captioning [22], [23], and visual question answering [24]. VSE approaches learn a joint visual-text space, e.g., via a triplet or contrastive loss, where the distance between embedded samples reflects their semantic relationship, and cross-modal attention [25], [26]. As a side experiment, we conducted a transformer-based pretraining involving both masked language modeling and image-text matching objectives to achieve better visual features. Still, progress in transformers is orthogonal to our primary aim as it does not consider how strong of a signal a caption provides for its co-occurring image.

There is also work to associate phrases in the caption to visually depicted objects [27] but none enable training of an independent object detector with accurate localization and classification, as we propose. In recent work, [28] predict masked words without localization, but use surrounding text at test time, unlike our models.

In Thomas and Kovashka [29], we show that image-text matching fails when the relation between an image and its corresponding (co-occurring) text is complementary rather than redundant. What this means for training object detection models from language supervision, is that the category overlap between image and co-occurring text may be low. To cope with this, our method exploits the structure of each unimodal space (image and text), and compares those structures, to compute how relevant each caption is for each image, and thus, whether the image-caption pair should be used for training object detectors.

2.3 Learning Visual Representations From Text

Recent work [30], [31], [32], [33], [34] aims to learn visual representations from their corresponding textual counterparts. Gomez *et al.* [30] predict the text LDA topic distribution from the image feature. Miech *et al.* [31] assume an MIL nature in videos, and use Noise Contrastive Estimation (NCE) to optimize the alignment between video clips and associated narrations. Desai and Johnson [33] harvest visual representations from training bidirectional captioning models and note the importance of predicting all caption tokens to learn a good visual representation. Radford *et al.* [34] optimize a classical co-attention model but learn the feature representation on a large dataset of 400 M image-text pairs. However, these methods do not train standalone object detectors. Bertasius *et al.* [32] apply a transformer-based language model to encode the text and match the visual feature extracted by an object detection model. They only optimize the representation to classify objects, while we also care about the detection scores and learn them in the unified framework since the visual proposals we use (Selective Search) are not as accurate as detection results. Most related to our work is Chen *et al.* [35]. The algorithm in this work discovers and localizes new objects from documentary *videos* by associating subtitles to video tracklets. They extract keywords from the subtitles using TFIDF. However, video provides benefits we cannot leverage, e.g., numerous frames containing nearly identical object instances. Importantly, we show that only using words that actually appear in the caption (as done in [35] with TFIDF) results in suboptimal performance compared to our method. Further, many components of Chen *et al.*'s method, e.g., the restriction to animal classes and the reliance on tracking, limit generalizability to other vocabularies and to images.

In our preliminary work, Ye *et al.* [36], we show we can successfully leverage unstructured supervision (highly descriptive captions well-aligned with the visual modality) but we do not explore any filtering or weighting of image-caption pairs. This weighting allows us to bypass the need for training a text-only classifier (which required a small amount of class labels), replacing it with techniques that require only image-caption pairs. In this work, we also include results on noisier datasets (MIRFlickr1M and Conceptual Captions) and without ImageNet pretraining.

2.4 Visual Reporting Bias and Concreteness

Our results show there is a gap between what humans name in captions, and what categorical annotations they provide. [37] study a similar phenomenon they refer to as "human reporting bias". They model the presence of an object as a latent variable, but we do the opposite—we model "what's in the image" by observing "what's worth saying". Further, we use the resultant model as precise supervision to guide detection model training.

Our work also measures how abstract is the connection between an image and a co-occurring text. Prior work predicts whether image and text that co-occur have a direct or complementary relationship [38], [39], e.g., whether the relation between image and its caption is "visible", "story", "subjective" or "meta" [39]. Unlike our method which also implicitly measures abstractness, these methods require additional annotations, aim for a discrete rather than continuous abstractness score, and are not applied in an object detection setting. Also related is work that measures how tightly clustered the visual companions of a word are [40] but this approach only computes scores for individual words, not for the relationships within image-caption pairs. In an auxiliary task, we compute the potential of [39], [40] to predict which captions may serve as clean supervision for weakly-supervised object detection, and we show that our method is equally or more promising.

3 Approach

We train object detectors from supervision only consisting of noisy captions and corresponding images. In realistic scenarios, captions and images may contain complementary information. We hypothesize that even for crowdsourced, descriptive captions which closely follow the image (e.g., COCO), not all caption-image pairs provide equally strong supervision, as some captions will overlap with the image to a stronger degree. Fig. 1 (top) shows two images, the first with high image-text alignment, where two objects (highlighted) are both shown and mentioned. The second image contains concepts that are visually not shown or are visually ambiguous (e.g., display, fabrics), hence extracting concrete nouns (objects) is more challenging. Thus, the first step in our framework (Section 3.1) is to determine which image-caption pairs to use as supervision: we propose two alternative approaches, one which uses a hard cutoff over the image-text alignment score, and another which uses all image-caption pairs but gives them different weight.

After selecting image-caption pairs for training, we next extract discrete labels at the image level (Section 3.2). We do so through a variety of techniques, the simplest of which is looking for exact string match between nouns in the caption and object words, and the most complex being training a classifier which takes in a caption (without a paired image) and maps this caption to a discrete set of labels (which may or may not be mentioned in the caption). Finally, given these pseudo ground-truth image-level labels, we train a variant of a prior weakly-supervised object detection technique: it first computes initial scores for each region and each object class, then refines these iteratively (Section 3.3).

3.1 Filtering Captions by Estimated Supervision Purity

We propose to filter image-caption pairs that are unlikely to be useful for training. The key idea is to estimate to what extent an image caption and the image provide overlapping (redundant) or complementary information. While complementarity is useful in general, for detection we require redundancy, i.e., the same objects being both shown and mentioned in the caption. We first describe an *example* of how to learn a joint image-text embedding; we do not require any particular technique for this part. We then compute homogeneity, i.e., how visually similar semantic neighbors of an image (images whose captions are similar in a *unimodal* word embedding space) are in the learned *joint, multimodal* embedding space. This homogeneity allows us to estimate the overlap of the image-caption pair. The computation of homogeneity follows our prior work [41], but was never used for object detection before. The process is overviewed in Fig. 2.

3.1.1 Preliminaries

Let $\mathcal{D} = \{\mathbf{I}, \mathbf{T}\}$ represent a dataset of n image-text pairs, where $\mathbf{I} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{T} = \{y_1, y_2, \ldots, y_n\}$ are the set of images and text (captions), respectively, and y_i is text co-occurring with image x_i (the two are semantically related). To reason about the relationships of images and text, we seek a joint manifold $\mathcal{M}$. For images, a convolutional network $f : \mathbf{I} \rightarrow \mathcal{M}$ is used to project images into the joint space, while a recurrent network $g : \mathbf{T} \rightarrow \mathcal{M}$ projects text. To obtain $\mathcal{M}$, we can use any cross-modal retrieval method. We describe two possibilities, triplet loss [42] and polysemous embedding model (PVSE) [21].

We first consider a simple triplet loss to derive $\mathcal{M}$

$$\mathcal{L}_{trip} = \Big[||x_i - y_i||_2^2 - ||x_i - y_j||_2^2 + m\Big]_+, \tag{1}$$

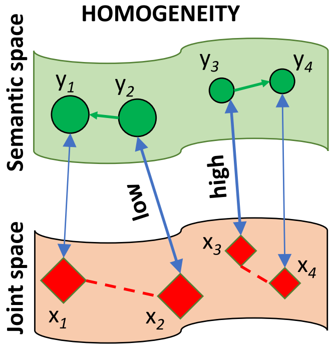


Fig. 2. Image-caption weighting: Green arrows connect neighbors in the original semantic space. Blue links show co-occurring images and text. Images whose texts are close in semantic space, which are close in the joint space (short red links), have high *homogeneity* scores [γ's in Eq. (9)].

where x_i, y_i appear together, while x_i, y_j do not, + denotes hinge loss, and m is a margin. Alternatively, PVSE uses a multiple-instance variant of triplet loss (K meanings hence K embeddings per sample), along with self-attention for visual and text features (not shown). The similarity of image x_i and text y_i is

$$s(x_i, y_i) = \max_{(k_1,k_2)\in\{1,\ldots,K\}\times\{1,\ldots,K\}} \left\langle \frac{x_{i_{k_1}}}{\|x_{i_{k_1}}\|_2}, \frac{y_{i_{k_2}}}{\|y_{i_{k_2}}\|_2} \right\rangle. \qquad (2)$$

3.1.2 Homogeneity

To capture how well-aligned an image and its corresponding caption are, we measure visual homogeneity (similarity) of the semantic concepts that an image illustrates. In other words, are images corresponding to semantically similar texts, visually similar? To measure this homogeneity, we first discover each image-text pair's *semantic neighbors* in text space $\Omega(\mathbf{T})$. Following [29], we compute neighbors in text space[2] because the text domain provides the cleanest semantic representation of the image-text pair. Let $\Psi(\Omega(y_i)) = \{\langle x'_{i_n}, y'_{i_n}\rangle\}_{n=1}^{N}$ represent the semantic nearest neighbor function over $\Omega(\mathbf{T})$, where $\{\langle x'_{i_n}, y'_{i_n}\rangle\}_{n=1}^{N}$ denotes the set of the N neighbors of $\langle x_i, y_i\rangle$ and $\langle x_i, y_i\rangle \notin \Psi(\Omega(y_i))$.

We next measure the homogeneity of the semantic neighbors in both the image and text domains, using the $f : \mathbf{I} \rightarrow \mathcal{M}$ and $g : \mathbf{T} \rightarrow \mathcal{M}$ projections of image and text *into the joint space*. Because our formulation is equivalent for both image/text neighbors, we let s_i represent a sample from either domain but require samples s_i and s_j come from the same domain. Let $\mathbf{s}'_i = [s'_{i_1}, s'_{i_2}, \ldots, s'_{i_N}]^{\intercal}$ denote the $N \times H$ matrix of embeddings of the neighbors of s_i found via Ψ, and $\mathbf{U} = \mathbf{s}'_i \mathbf{s}'^{\intercal}_i$ compute the pairwise similarities between all semantic neighbors through cross-product. We compute the *homogeneity score* α_i^{HOM} for s_i as follows:

$$\alpha_i^{HOM} = \frac{1}{N^2} \sum_{r=1}^{N} \sum_{c=1}^{N} \mathbf{U}_{(r,c)}, \qquad (3)$$

where r, c index over the rows and columns of $\mathbf{U} = \mathbf{s}'_i \mathbf{s}'^{\intercal}_i$. For the different image ($s_i = x_i$) and text ($s_i = y_i$) domains, we compute *visual homogeneity score* $\boldsymbol{\alpha}_{(I)}^{HOM}$ and *text homogeneity score* $\boldsymbol{\alpha}_{(T)}^{HOM}$, respectively. Both of these scores capture how aligned an image and its co-occurring text are; thus, *higher* $\boldsymbol{\alpha}$ *scores indicate image-text captions from which supervision signal can more reliably be extracted*. We also consider the difference of $\boldsymbol{\alpha}_{(I)}^{HOM}$ and $\boldsymbol{\alpha}_{(T)}^{HOM}$ as an indicator for supervision purity.

Note that the cost for computing homogeneity score is neglectable in that we *offline preprocessed* all examples in the training set by caching sample embeddings into a memory bank. We only find semantic neighbors once using a pretrained Doc2Vec. Then, computing $\boldsymbol{\alpha}^{HOM}$ weights is efficient as it only requires multiplication.

3.1.3 Scoring Captions for Homogeneity

For object detection, we prefer to train from examples in which objects are both shown in the image and mentioned in the text. Section 3.1.2 provides a way to measure the redundancy between the image and text modalities. Here, we describe how to use this measure.

Filtering. We hypothesize that selecting training data and filtering out noisy image-caption pairs using $\boldsymbol{\alpha}_{(I)}^{HOM}$ and $\boldsymbol{\alpha}_{(T)}^{HOM}$ will improve the detection model training. We provide an experiment in Section 4.4, which selects the 30,000 image-caption pairs from COCO that have the highest homogeneity scores. Our method provides significantly better detection results than random selection (Table 5).

Weighting. Hard-cutoff filtering requires finding the right cutoff value (e.g., top-30k), and it means discarding some potentially useful data. Compared to the filtering strategy, weighting does not require a hard cutoff and is more data-efficient. It applies different weights to image-caption pairs. For image-caption pairs that are more overlapped (i.e., high homogeneity), weighting assigns large weight to the loss term in that these examples will likely be useful for training detection models. For image-caption pairs that are more complementary, weighting assigns small weights because the information may not well-aligned. In Eq. (9), we use $\boldsymbol{\alpha}_{(I)}^{HOM}$ as the heuristic weighting factor γ. We provide an ablation in Section 4.4, and Tables 6, 7 and 8 show the impact of using the weighting factor $\boldsymbol{\alpha}_{(I)}^{HOM}$.

The strategy described in this section allows us to select captions with useful supervision, thus can be coupled with simple mechanisms to extract labels at the image level from captions, e.g., ExactMatch in the next subsection. We do not expect it to improve results when a mapping function from captions to training labels is learned with supervision.

3.2 Pseudo Training Label Inference From Text

After getting the image-caption pairs estimated to be well-aligned, we now proceed to extract pseudo object labels from the selected noisy captions, to benefit weakly-supervised object detection. The foundation of WSOD builds on an important assumption from MIL (Eq. (7)), which suggests that *precise* image-level labels should be provided. The straightforward solution is to extract object labels from captions via lexical matching. However, it has limitations. Consider an image with three sentence descriptions:

"a ***person*** *is riding a* ***bicycle*** *on the side of a bridge."*
"a ***man*** *is crossing the street with his* ***bike****."*
"a ***bicyclist*** *peddling down a busy city street."*

Only the first sentence exactly matches the categories "person" and "bicycle". Even if we allow synonyms of "man" and "person" or "bicycle" and "bike", only the first two precisely describe both objects, while the last one still misses the instance of "bicycle" unintentionally. When using these examples to train object detectors, the first two instances may bring positive effect, but the last one will be wastefully discarded as false negative i.e., not relevant to the categories "person" or "bicycle". As further examples, in Fig. 1 (bottom-right), none of the captions (one shown) mention the "bowls" or "spoons" that are present, and only some mention "oven". Finally, in Fig. 6, the caption mentions a "suit" worn by a speaker at a conference, but not the "tie", even though one is present.

This observation inspires us to amplify the supervision signal that captions provide, and squeeze more information

[2] Specifically, Doc2Vec [43] due to its appropriateness for longer texts, although BERT [44] could also be used.

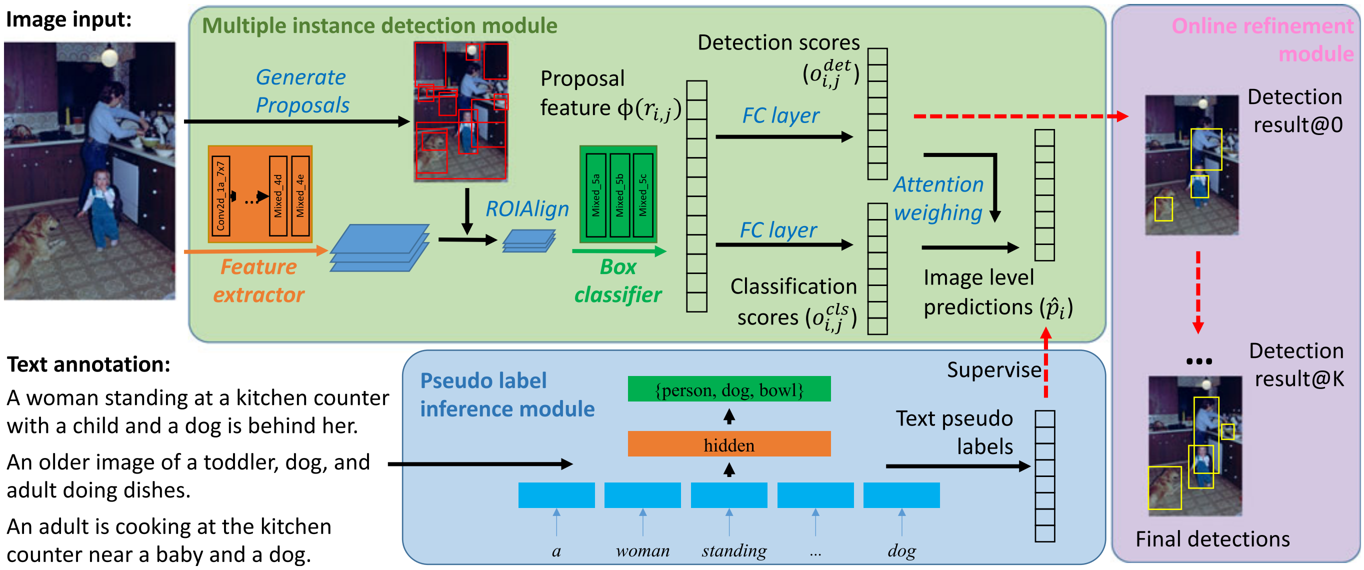


Image input:
Multiple instance detection module
Generate Proposals
Proposal feature $\phi(r_{i,j})$
Detection scores $(o^{det}_{i,j})$
FC layer
Attention weighing
FC layer
Conv2d_1a_7x7
Mixed_4d
Mixed_4e
Feature extractor
ROIAlign
Mixed_5a
Mixed_5b
Mixed_5c
Box classifier
Classification scores $(o^{cls}_{i,j})$
Image level predictions $(\hat{p}_i)$
Online refinement module
Detection result@0
Detection result@K
Final detections
Text annotation:
A woman standing at a kitchen counter with a child and a dog is behind her.
An older image of a toddler, dog, and adult doing dishes.
An adult is cooking at the kitchen counter near a baby and a dog.
Pseudo label inference module
{person, dog, bowl}
hidden
a
woman
standing
...
dog
Supervise
Text pseudo labels

Fig. 3. Harvesting detection models from free-form text. We propose to use a pseudo training label inference module (bottom) to amplify signals in free-form texts to supervise the learning of the multiple instance detection network (top). The detection model is refined by an online refinement module (right) to produce the final detection results. Detection ($o^{det}_{i,j} \in \mathbb{R}^C$) and classification ($o^{cls}_{i,j} \in \mathbb{R}^C$) scores and image predictions $\hat{p}_i \in \mathbb{R}^C$ refer to predictions for all classes.

out of them. Fig. 3 (bottom) shows the approach we use to amplify the signal. This text-only model takes free-form texts as input, embeds individual words to a 300D space using GloVe [45], and projects the embedded features to a 400D latent space. We then use max-pooling to aggregate the word-level representations. Then, we use this intermediate representation to predict the implied instances (e.g., 80 classes as defined in COCO, or any other categories); this prediction answers "what's in the image" and serves as pseudo image-level training labels, to be used in Section 3.3.

There exists a subtle balance when using pseudo labels to train object detectors. Admittedly, our strategy increases the recall rates thus more data could be utilized. However, with the increased recall, precision will drop inevitably thus the fundamental assumption in MIL is threatened. Specifically, the *precise label* assumption makes the model very sensitive to false positive cases: when inappropriate labels are given where none of the proposals have a good response, the model gets confused, resulting in non-optimal detections. Thus, we adopt a two-step procedure: first we look for an exact match of object labels from captions, following the intuition that *explicitly mentioned objects* should be significant and obvious enough in the image; second, when no object can be matched, we use our label inference model to predict labels as *unspoken intended objects* to guide the object detection. We show our method TEXTCLSF outperforms several strong alternatives that also infer pseudo labels.

Alternative Strategies. We also experiment with alternative multiple pseudo-label generation techniques when lexical matching (EXACTMATCH) fails to find a match. First, we consider a *manually constructed, hence expensive* COCO synonym vocabulary list (EXTENDVOCAB) which maps 413 words to 80 categories [46]. Another variant, GLOVE, takes advantage of GloVe word embeddings [47], assigning pseudo-labels for a sentence by looking for the category that has the smallest embedding distance to any word in the sentence. We also finetune the GloVe word embeddings on COCO using a visual-text ranking loss, and use the pseudo labels retrieved by the resultant text embedding, resulting in LEARNEDGLOVE.

Discussion. Our text classifier relies on both captions and category labels. Once the bridge between captions and labels is established, it generalizes to other datasets, as we show in Table 1. Importantly, we only need a small fraction of labels to train this text classifier; as we show in Fig. 5, precision has a small range (between 89% and 92%) when we use between only 5% and 100% of the COCO data, while recall is stable at 62%. Thus, our text model could learn from a *single source* dataset with *a few* labels, then it could transfer the knowledge to other *target* datasets, requiring only free-form text as supervision. If no labels are available, the caption weighting strategy (Section 3.1) can be paired with EXACTMATCH or other alternatives to TEXTCLSF. However, TEXTCLSF performs slightly better overall.

3.3 Detection From Inferred Pseudo Image Labels

We next describe how we use the inferred pseudo labels at the image level, to train an object detection model. As shown in Fig. 3 (top), we first extract proposals with accompanying features. An image is fed into randomly initialized or pretrained (on ImageNet [47]) convolutional layers. Then, *ROIAlign* is used for cropping the proposals (at most 500 boxes per image) generated by *Selective Search* [48], resulting in fixed-sized convolutional feature maps. Finally, a box feature extractor is applied. If $\{r_{i,j}\}_{j=1}^{m}$ are the m proposals of a given image x_i, this process results in proposal feature vectors $\{\phi(r_{i,j})\}_{j=1}^{m}$ where each $\phi(r_{i,j}) \in \mathbb{R}^d$. Note that even when our model is pretrained on ImageNet, it *does not leverage* any image labels on the datasets on which we train and evaluate our detection models (PASCAL and COCO).

3.3.1 *Initial Detection Scores*

We next introduce the prediction of image-level labels $\hat{p}_{i,c}$ ($c \in \{1, \ldots, C\}$ for the i-th image, where C is the number of classes) and of detection scores. If not noted otherwise, we use i to index training examples, j to index region proposals within an image, and c to index class labels. The method described in this section follows prior work, i.e., Bilen *et al.* [6], but labels used for training are potentially noisy as they come from our pseudo label inference module, Section 3.2.

First, we feed the proposal features $\phi(r_{i,j})$ into two parallel fully-connected layers to compute the detection scores $o^{\text{det}}_{i,j,c} \in \mathbb{R}^1$ (top branch in the green MIL module in Fig. 3) and classification scores $o^{\text{cls}}_{i,j,c} \in \mathbb{R}^1$ (bottom branch), in which both scores are related to a specific class c and the j-th proposal of image x_i

$$o^{\text{cls}}_{i,j,c} = w_c^{\text{cls}\intercal}\phi(r_{i,j}) + b_c^{\text{cls}}, \qquad o^{\text{det}}_{i,j,c} = w_c^{\text{det}\intercal}\phi(r_{i,j}) + b_c^{\text{det}}. \tag{4}$$

We convert these scores into: (1) $p^{\text{cls}}_{i,j,c}$ the probability that object c presents in the j-th proposal; and (2) $p^{\text{det}}_{i,j,c}$ the probability that the j-th proposal is important for predicting image-level label $y_{i,c}$

$$p^{\text{cls}}_{i,j,c} = \sigma(o^{\text{cls}}_{i,j,c}), \qquad p^{\text{det}}_{i,j,c} = \frac{\exp(o^{\text{det}}_{i,j,c})}{\sum_{j=1}^{m}\exp(o^{\text{det}}_{i,j,c})}. \tag{5}$$

Finally, the aggregated image-level prediction is computed as follows, where greater values of $\hat{p}_{i,c} \in [0,1]$ mean higher likelihood that c is present in the image x_i

$$\hat{p}_{i,c} = \sigma\left(\sum_{j=1}^{m} p^{\text{det}}_{i,j,c}\, o^{\text{cls}}_{i,j,c}\right). \tag{6}$$

Assuming the label $y_{i,c} = 1$ if and only if class c is present in the input image x_i, the **m**ultiple **i**nstance **d**etection loss used for training the model is defined as

$$L_{\text{mid}}(x_i, y_i) = -\sum_{c=1}^{C}\left[y_{i,c}\log\hat{p}_{i,c} + (1 - y_{i,c})\log(1 - \hat{p}_{i,c})\right]. \tag{7}$$

The weakly supervised detection score given both proposal $r_{i,j}$ and class c is the product of $p^{\text{cls}}_{i,j,c}$ and $p^{\text{det}}_{i,j,c}$ which is further refined as described in Section 3.3.2.

3.3.2 *Online Instance Classifier Refinement*

The third component of our WSOD model is Online Instance Classifier Refinement (OICR), as proposed by Tang *et al.* [8]. Given a ground-truth class label, the top-scoring proposal, as well as proposals highly overlapping with it, are selected as references. These proposals are treated as positives for training the box classifier of this class while others are treated as negatives. The initial top-scoring proposal may only partially cover the object, so allowing highly-overlapped proposals to be treated as positives gives them a second chance to be considered as containing an object, in the subsequent model refinement. This reduces the chance of propagating incorrect predictions. In addition, sharing the convolutional features between the original and refining models makes training more robust.

Following [8], we stack multiple refining classifiers and use the output of the previous one to generate instance-level supervision to train the successor. The detection score at the 0-th iteration is computed using $s^{(0)}_{i,j,c} = p^{cls}_{i,j,c}\ p^{det}_{i,j,c} (c \in \{1,\ldots,C\}), s^{(0)}_{i,j,C+1} = 0$ (where $C+1$ is the background class). Given the detection score $s^{(k)}_{i,j,c}$ at the k-th iteration, we use the image-level label to get the *instance-level* supervision $y^{(k+1)}_{i,j,c}$ at the $(k+1)$-th iteration. Assume that c' is a label attached to image x_i, we first look for the top-scoring box $r_{i,j'}$ ($j' = \text{argmax}_j s^{(k)}_{i,j,c'}$) responsible to predict c'. We then let $y^{(k+1)}_{i,j,c'} = 1, \forall j \in \{j | IoU(r_{i,j}, r_{i,j'}) > threshold\}$. When $k > 0$, $s^{(k)}_{i,j,c}$ is inferred using a $(C+1)$-way FC layer, as in Eq. (4). The OICR training loss is defined in Eq. (8)

$$L^k_{\text{oicr}}(x_i, y_i) = -\frac{1}{m}\sum_{j=1}^{m}\sum_{c=1}^{C+1}\hat{y}^{(k)}_{i,j,c}\log s^{(k)}_{i,j,c}, \qquad k = 1,\ldots,K. \tag{8}$$

Unlike the original OICR, our WSOD module aggregates logits instead of probability scores, which in our experience stabilizes training. We also removed the reweighing of untrustworthy signals emphasized in [8] since we found it did not contribute significantly.

The final loss we optimize is Eq. (9), where γ_i is the *per-example* weighting factor. $\gamma_i = 1$ for all i if we are not applying homogeneity weighting. If we use hard filtering based on $\alpha^{HOM}_{(I)\ i}, \alpha^{HOM}_{(T)\ i}$, then $\gamma_i = 1$ for samples included in training, and $\gamma_i = 0$ for others. If using image-caption weighting, $\gamma_i = \alpha^{HOM}_{(I)\ i}$. We refine our model 3 times ($K = 3$) if not mentioned otherwise

$$L(x_i, y_i) = \gamma_i\left(L_{\text{mid}}(x_i, y_i) + \sum_{k=1}^{K} L^k_{\text{oicr}}(x_i, y_i)\right). \tag{9}$$

3.4 Implementation Details

We first obtain a joint image-caption embedding space, by training a triplet loss model [42] (with margin m set to 0.5) or PVSE [21] model (with $K = 3$ embeddings per sample and margin 0.1, using the COCO validation set). We use this model to infer image-text joint embeddings. Then we use Eq. (3) to infer the homogeneity scores. We use ResNet-50 [49] initialized randomly or with ImageNet features for images (both types of results shown in Table 8). Images are scaled to 224 x 224 and augmented with random horizontal flipping. For text, we use GRU [50] with hidden state size 512, initialized with 200D word embeddings learned on the COCO dataset, using [51]'s implementation of Doc2Vec, distributed memory [43], 20 epochs with window size of 20, and ignoring words that appear less than 20 times. We use Xavier initialization [52], the Adam optimizer [53] with minibatch size of 32, learning rate 1.0e-4 (decayed by 10x after every 5 epochs of no decrease in val loss), and weight decay 1e-5. We use [54] to efficiently compute approximate nearest neighbors for Ψ and use $N = 200$ nearest neighbors.

For the text classifier which predicts the pseudo image-level labels, we adopt a multi-layer perceptron (see Fig. 3 bottom). We first embed word tokens to 300D GloVe embeddings and project them to a 400D latent space. We use max-pooling to aggregate these word-level representations to get the fixed-length 400D caption representation. Next, we use this max-pooled intermediate representation to predict the implied instances (e.g., 80 classes in COCO). The object labels are used to supervise the text classifier learning (crossentropy loss). We use the AdaGrad optimizer [55], learning rate of 0.1, and batch size of 20.

To train the weakly supervised object detection model, we first use Selective Search [48] from OpenCV [56] to extract at most 500 proposals for each image. We follow the "Selective search quality" parameter settings in [48]. We use Selective Search because it is a generic, dataset-independent proposal generation procedure, as opposed to other CNN-based alternatives which are trained end-to-end from a specific dataset in a supervised fashion. We use TensorFlow [57] as our training framework. To compute the proposal feature vectors, we use the layers ("Conv2d_1a_7x7" to "Mixed_4e") from Inception-V2 [10] to get the conv feature map, and the layers ("Mixed_5a" to "Mixed_5c") from the same model to extract the proposal feature vectors after the ROIAlign operation. The Inception-V2 model is either randomly initialized, or pretrained on ImageNet [47]; the supervised detector counterpart of our model, using this architecture, was explored by [58]. To augment the training data, we resize the image randomly to one of the four scales $\{400, 600, 800, 1200\}$. We also randomly flip the image left to right at training time. At test time, we average the proposal scores from the different resolution inputs. We set the number of refinements to 3 for the OICR since it gives the best performance. For post-processing, we use non-maximum-suppression with IoU threshold of 0.4. We use the AdaGrad optimizer [55], a learning rate of 0.01, and a batch size of 2 as commonly used in WSOD methods [8], [16]. The models are usually trained for 100K iterations on Pascal VOC (roughly 40 epochs on VOC2007 and 17 epochs on VOC2012) and 500K on COCO (8.5 epochs), using a validation set to pick the best model. Our implementation is available at `https://github.com/yekeren/Cap2Det`.

4 Experiments and Results

First, we present our experimental settings (Section 4.1). Second, we compare the accuracy of alternative strategies on *pseudo image-level training label extraction* from captions (Section 4.2). Third, we show that our approach achieves strong *detection* performance using supervision from captions (Section 4.3). By training on COCO captions, we achieve close to state-of-the-art results on weakly supervised detection on PASCAL, even though the supervision we leverage is weaker than competitor methods. Importantly, our text classifier trained on COCO generalizes to other datasets, and allows us to use Flickr30K and the noisier datasets MIRFlickr1M and Conceptual Captions which do not feature clean, descriptive captions. In all settings, our primary pseudo label inference method, EM+TextClsf, outperforms the alternative techniques, including the ExactMatch baseline. Fourth, we show the improvements achieved by filtering and scoring noisy image-caption examples (Section 4.4). We conclude that the redundancy between image and text is key to train a successful weakly-supervised detection model. Finally, as a sanity check, we show our approach performs competitively to prior methods on the task of learning from *clean, ground-truth* image-level labels (Section 4.5).

Note that in all experiments, we focus on evaluating the impact of a single component of our model, focusing on the caption weighting and pseudo label inference. For most experiments, we determine success by comparing to an upper bound (e.g., ground-truth labels) and/or a lower bound (e.g., naive training label extraction, unweighted loss using captions equally). We show our methods' advantages persist when we replace Section 3.3 with alternative state-of-the-art techniques, e.g., Ren *et al.* [9].

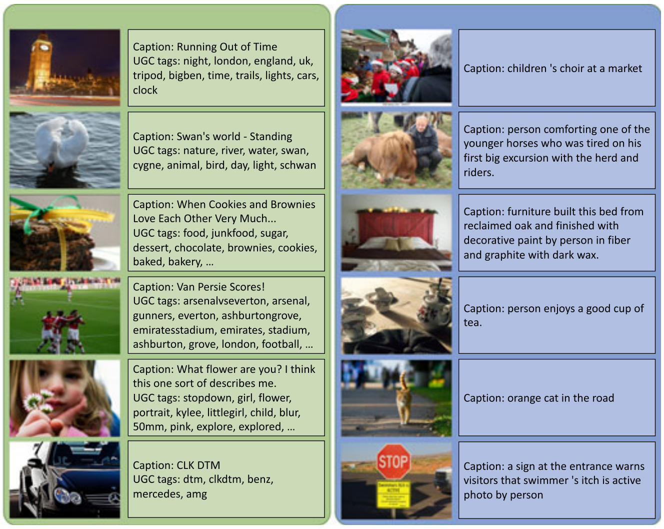
4.1 Datasets and Metrics

Our experiments involve the datasets COCO, PASCAL VOC, Flickr30K, MIRFlickr1M, and Conceptual Captions.

PASCAL Visual Object Classes (VOC) [59] is a standard image dataset for object class recognition. It focuses on a limited number of classes (20 objects). We use it to evaluate our learned object detection models (4,952 and 10,991 test examples in VOC07 and VOC12, respectively).

Common Objects in Context (COCO) [60] is a large-scale object detection, segmentation, and captioning dataset. We use: (1) its image-caption data to train the triplet and PVSE models necessary to infer image-caption homogeneity (Section 3.1); (2) its caption-label data to train our pseudo label inference module (Section 3.2) and test it on three other datasets; (3) its 118K training images, each paired with 5 captions, to train our detection model (Section 3.3) using 591,435 COCO [61] captions paired to the 118,287 *train2017* images. ExactMatch fails to extract any label for roughly 15K *train2017* instances. Since COCO is fully annotated with instance-level boxes, we use its evaluation server to measure performance of the resulting object detection models.

Flickr30K and MIRFlickr1M. We use the Flickr data to prove that our weakly-supervised object detection models can generalize to alternative datasets. Flickr30K [62] (30K images, each paired with 5 captions) contains crowdsourced captions. However, we also use 200K examples from the 1M noisy data in MIRFlickr1M [63] (subset for computational reasons). The dataset pairs images with user-generated content, which is not guaranteed to describe the content of the images in an exhaustive or precise manner. ExactMatch


Caption: Running Out of Time
UGC tags: night, london, england, uk, tripod, bigben, time, trails, lights, cars, clock

Caption: Swan's world - Standing
UGC tags: nature, river, water, swan, cygne, animal, bird, day, light, schwan

Caption: When Cookies and Brownies Love Each Other Very Much...
UGC tags: food, junkfood, sugar, dessert, chocolate, brownies, cookies, baked, bakery, ...

Caption: Van Persie Scores!
UGC tags: arsenalvseverton, arsenal, gunners, everton, ashburtongrove, emiratesstadium, emirates, stadium, ashburton, grove, london, football, ...

Caption: What flower are you? I think this one sort of describes me.
UGC tags: stopdown, girl, flower, portrait, kylee, littlegirl, child, blur, 50mm, pink, explore, explored, ...

Caption: CLK DTM
UGC tags: dtm, clkdtm, benz, mercedes, amg

Caption: children 's choir at a market

Caption: person comforting one of the younger horses who was tired on his first big excursion with the herd and riders.

Caption: furniture built this bed from reclaimed oak and finished with decorative paint by person in fiber and graphite with dark wax.

Caption: person enjoys a good cup of tea.

Caption: orange cat in the road

Caption: a sign at the entrance warns visitors that swimmer 's itch is active photo by person

Fig. 4. Example image-caption pairs from MIRFlickr1M and Conceptual Caption datasets. For MIRFlickr1M, captions and tags are written by the uploader and website users. For ConcCap, captions are parsed from the alt-text HTML attribute associated with web images. Both datasets did not crowdsource i.e., pay workers to label the images. Note how often "person" is mentioned on the right but not visible.

fails to extract labels on roughly 113K of those. We use captions from these datasets to train our detection models. Fig. 4 shows examples.

Conceptual Captions [64] contains 3.3M images annotated with captions. The captions are from the web, and illustrate the noisy web data environment. We conduct an experiment that tries to benefit from the large corpus via vision-language pretraining. We also directly train object detectors on Conceptual Captions. To keep results comparable to those when using COCO Captions, we extract a 118K subset from Conceptual Captions. ExactMatch fails to extract any label on roughly 54K samples in that subset. Conceptual Captions includes alt-text for images. Alt-texts are preprocessed with hypernimization to replace named entities (e.g., architect's name) with object names (e.g., "person"). As can be seen from Fig. 4 (right), this alt-text is not descriptive of the images in the same way that COCO captions are descriptive. For example, the third caption mentions "person", but this person likely refers to whoever made the decorative paint; *this person is not visible in the image.* Similarly, the "person enjoying tea" (fourth caption) is also not visible. Conversely, the persons in the first example are not explicitly mentioned, and neither is the car in the bottom example. This makes straightforward techniques for extracting labels (e.g., ExactMatch) prone to failure due to missed or incorrect labels.

Evaluation Protocols. We follow the standard protocols used in VOC and COCO to fairly compare to other methods. On VOC, we report the mean Average Precision (AP) at IoU >0.5 and we also report the per-class AP. On COCO, we report the same AP@IoU0.5 as in VOC, and the mean Average Precision across different IoU thresholds, i.e., mAP@IoU.5:.05:.95. We also include the full metrics in COCO such as AP of differently-sized objects. We submit our results to the VOC and COCO evaluation servers to get the VOC07 and COCO testing results.

4.2 Accuracy of Pseudo Image Label Inference

We first test the generalizing power of our pseudo image-level label inference module (Section 3.2). Our results show that without using too much data, one can train a reasonably good text classifier inferring accurate image-level labels from paired captions.

In Fig. 5 we show the *precision* and *recall* of these label inference methods evaluated directly on the COCO image-level labels (5,000 examples of the *val2017* set). We observe that EM+ExtendVocab, which uses the hand-crafted word-synonyms dictionary, provides the best recall (60.6%) among all methods but the worst precision of 81.1%. The word-embedding-based top-scoring matching methods of EM+GloVe and EM+LearnedGloVe provide precise predictions (84.5% and 84.7% respectively, which are the highest). However, our EM+TextClsf achieves significantly improved precision compared to these. We would like to point out that while in Tables 1 and 2, our method uses the full COCO training set (118,287 concatenated captions), it achieves very similar performance with even a small fraction of the data. With 5% of the data (6,000 caption-label pairs), the method achieves 89% precision (versus 92% precision with 100% of the data), both of which are much higher than any other baselines; recall is about 62% for both 5% and 100% training data. Thus, it is sufficient to use a small portion of precise text labels to train a generalizable label inference classifier, and the knowledge can transfer to other datasets as we show in Table 1.

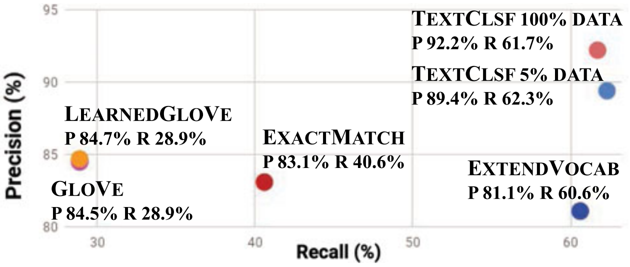


Fig. 5. Analysis of different text supervision. We compare the pseudo labels (Section 3.2) to COCO *val* ground-truth.

To better understand the generated labels, we show two qualitative examples in Fig. 6. The image on the right shows that our model infers "tie" from the observation of "presenter", "conference" and "suit", while all other methods fail to extract this object category for visual detection.

It is also interesting to measure this performance per class, as we show in Fig. 7. The lexical matching method ExactMatch performs similarly to EM+TextClsf in terms of precision (not shown). For both ExactMatch and EM+TextClsf, recall is very low for bottle, car, and chair, indicating these are common objects which however are *usually not mentioned in captions*. In contrast, other common objects (e.g., cat) have high recall because they are usually mentioned when present in the image. However, for classes such as boat, cow, and person, ExactMatch *has much lower recall rate than* EM+TextClsf. We thus quantitatively explain why EM+TextClsf is better than ExactMatch for these classes (Table 1 B, boat 29.9% versus 25.9% or 13.3% versus 9.6%; cow 61.2% versus 49.0% or 47.4% versus 28.0%; person 16.9% versus 10.4% or 10.7% versus 4.0%). By explicitly handling the noise in the captions (the lack of mentions of objects that do appear), we cope with the "human reporting bias" [37].

4.3 Comparing Label Inference Strategies for Detection

We compare our pseudo label inference module (Section 3.2) with alternative strategies to extract object labels from captions. All of these strategies are upper-bounded in terms of performance by using ground-truth image-level labels GT-Label. Note that apart from the strategy used to mine image-level labels for training, these strategies all use the same architecture and WSOD approach as our method (Section 3.3). We show combinations of the exact match strategy with these methods (when exact match fails), resulting in EM+GloVe, EM+LearnedGloVe, EM+ExtendVocab and EM+TextClsf. We examine how well these strategies leverage captions from COCO, Flickr30K, MIRFlickr1M, and Conceptual Captions for detection.

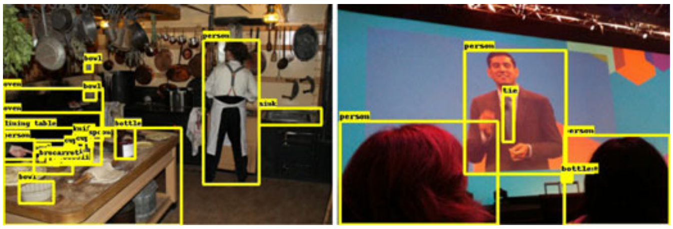
Training With COCO Captions. Table 1, segments (A) and (B), show the results on PASCAL VOC 2007. At the top (A) are two upper-bound methods that train on *ground-truth* image-level *labels*, while methods in (B) train on labels

TABLE 1
Average Precision (in %) on the VOC 2007 Test Set (Learning From COCO, Flickr30K, MIRFlickr1M, and Conceptual Captions Captions)

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
(A) Training on different datasets using ground-truth labels:																					
GT-Label VOC	68.7	49.7	53.3	27.6	14.1	64.3	58.1	76.0	23.6	59.8	50.7	57.4	48.1	63.0	15.5	18.4	49.7	55.0	48.4	67.8	48.5
GT-Label COCO	65.3	50.3	53.2	25.3	16.2	68.0	54.8	65.5	20.7	62.5	51.6	45.6	48.6	62.3	7.2	24.6	49.6	34.6	51.1	69.3	46.3
(B) Training on COCO dataset using captions:																					
ExactMatch (EM)	63.0	**50.3**	50.7	25.9	**14.1**	64.5	50.8	33.4	17.2	49.0	48.2	46.7	44.2	59.2	10.4	14.3	49.8	37.7	21.5	47.6	39.9
EM + GloVe	**66.6**	43.7	53.3	29.4	13.6	65.3	**51.6**	33.7	15.6	50.7	46.6	45.4	47.6	**62.1**	8.0	**15.7**	48.6	46.3	30.6	36.4	40.5
EM + LearnedGloVe	64.1	49.9	**58.6**	24.9	13.2	**66.9**	49.2	26.9	13.1	57.7	**52.8**	42.6	**53.2**	58.6	14.3	15.0	45.2	50.3	34.1	43.5	41.7
EM + ExtendVocab	65.0	44.9	49.2	**30.6**	13.6	64.1	50.8	28.0	**17.8**	59.8	45.5	**56.1**	49.4	59.1	16.8	15.2	**51.1**	**57.8**	14.0	**61.8**	42.5
EM + TextClsf	63.8	42.6	50.4	29.9	12.1	61.2	46.1	**41.6**	16.6	**61.2**	48.3	55.1	51.5	59.7	**16.9**	15.2	50.5	53.2	**38.2**	48.2	**43.1**
(C) Training on COCO dataset without ImageNet pretraining:																					
GT-Label COCO	40.7	17.2	5.7	0.8	0.0	31.9	21.4	30.6	0.3	31.0	26.0	30.5	23.9	43.8	13.8	7.0	25.5	2.7	15.3	34.9	20.15
ExactMatch (EM)	43.8	13.9	6.0	0.8	0.0	**39.6**	31.0	**34.2**	0.3	20.6	**28.8**	25.6	25.2	35.9	4.9	**3.3**	27.3	3.0	9.9	15.1	18.46
EM + GloVe	44.6	13.4	**9.1**	1.4	0.1	36.3	29.7	30.7	0.2	17.2	21.2	**29.3**	25.1	36.9	5.9	2.6	6.7	2.7	8.7	16.8	16.93
EM + ExtendVocab	**44.9**	12.8	8.3	1.1	0.0	35.5	**31.6**	27.0	**0.4**	**25.7**	27.1	27.1	22.0	34.8	**10.3**	0.2	**29.6**	**4.1**	**11.7**	**26.8**	19.05
EM + TextClsf	44.5	**18.4**	6.7	**1.5**	**0.3**	35.6	27.6	32.2	0.2	**25.7**	28.0	28.3	**26.0**	**38.8**	6.2	0.1	28.4	2.7	9	25.8	**19.30**
(D) Training on Flickr30K dataset using captions:																					
ExactMatch (EM)	**46.6**	**42.9**	42.0	9.6	7.7	31.6	44.8	53.2	13.1	28.0	39.1	43.2	31.9	**52.5**	4.0	**5.1**	38.0	28.7	**15.8**	41.1	31.0
EM + ExtendVocab	37.8	37.6	35.5	11.0	**10.3**	18.0	47.9	51.3	**17.7**	25.5	37.0	47.9	**35.2**	46.1	**15.2**	0.8	27.8	35.6	5.8	42.0	29.3
EM + TextClsf	24.1	38.8	**44.5**	**13.3**	6.2	**38.9**	**49.9**	**60.4**	12.4	**47.4**	**39.2**	**59.3**	34.8	48.1	10.7	0.3	**42.4**	**39.4**	14.1	**47.3**	**33.6**
(E) Training on Flickr30K dataset using captions, without ImageNet pretraining:																					
ExactMatch (EM)	1.4	14.5	2.4	**0.7**	0.0	**9.4**	19.2	3.5	0.2	**3.5**	9.2	**11.0**	10.0	23.0	1.6	0.0	0.4	2.1	2.8	2.8	5.9
EM + ExtendVocab	0.9	15.0	**2.9**	0.6	0.0	1.0	11.2	4.0	0.2	0.3	13.0	10.8	**10.9**	21.6	2.7	**0.1**	0.6	**3.9**	0.8	**6.6**	5.4
EM + TextClsf	**1.8**	**21.0**	1.9	**0.7**	0.0	5.4	**20.6**	**5.7**	0.2	2.1	**12.4**	**11.0**	10.4	**24.7**	**3.4**	**0.1**	**1.0**	1.4	**3.7**	4.1	**6.6**
(F) Training on MIRFlickr1M (200k subset) using captions, without ImageNet pretraining:																					
ExactMatch (EM)	7.6	7.5	8.5	0.5	0.1	13.0	18.9	15.2	0.4	2.9	2.6	6.6	3.7	21.9	3.3	0.2	2.4	**2.3**	1.4	16.2	6.8
EM + TextClsf	**31.7**	**10.5**	**8.9**	**0.8**	**0.3**	**15.5**	**28.9**	**28.1**	**0.6**	**6.3**	**6.5**	**20.5**	**12.9**	**26.6**	**15.1**	**0.6**	**4.1**	1.0	**3.5**	**19.1**	**12.1**
(G) Training on Conceptual Captions (118k subset):																					
ExactMatch (EM)	26.1	21.2	17.0	10.0	7.7	19.1	**31.1**	15.2	**7.0**	22.3	**35.9**	32.6	13.1	25.5	3.5	**4.6**	34.1	12.0	11.8	12.2	18.1
EM + TextClsf	**60.9**	**36.5**	**35.7**	**21.4**	**8.9**	**28.3**	20.1	**26.2**	4.4	**34.8**	17.5	**41.7**	**22.8**	**51.2**	**11.3**	0.3	**42.4**	**29.3**	**21.3**	**28.4**	**27.2**
(H) Training on Conceptual Captions (118k subset), without ImageNet pretraining:																					
ExactMatch (EM)	1.1	1.6	**0.7**	0.3	0.0	3.1	**10.4**	**14.2**	0.2	0.3	1.2	7.0	0.9	2.7	1.8	0.1	**0.5**	**0.7**	0.6	**0.2**	2.38
EM + TextClsf	**1.4**	1.6	0.5	0.3	0.0	**3.9**	7.0	13.5	0.2	**0.4**	**3.4**	**8.8**	**1.7**	**8.9**	**2.9**	0.1	0.3	0.6	**4.1**	0.1	**3.02**

We evaluate on only the overlapping 20 VOC objects. Best method per column (except GT methods) in bold. *Our proposed EM+TextClsf achieves 93-96% of GT-Label COCO, and exceeds ExactMatch (EM) by 8% with COCO captions and 27-78% with noisy captions (MIRFlickr1M and Conceptual Captions).*

extracted from image-level *captions*. ExactMatch (EM) performs the worst probably due to its low data utilization rate, as evidenced by the fact that all methods incorporating pseudo labels improve performance notably. Specifically, EM+GloVe uses knowledge of the pre-trained GloVe embeddings. It alleviates the synonyms problem to a certain extent, thus it improves the mAP by 2% compared to ExactMatch. However, the GloVe embedding is not optimized for the specific visual-captions, resulting in noisy knowledge transformation. EM+LearnedGloVe learns dataset-specific word embeddings. Its performance, as expected, is 3% better than EM+GloVe. The strongest baseline is EM+ExtendVocab, as the manually picked vocabulary list covers most frequent occurrences. However, collecting such vocabulary requires human effort, and is not a scalable and transferable strategy. Our EM+TextClsf outperforms this expensive baseline, especially for categories "cat", "cow", "horse", and "train". Finally, despite the noisy supervision, our EM +TextClsf almost bridges the gap to the GT-Label COCO upper bound in Table 1 (A).



A man is in a kitchen making pizzas .
Man in apron standing on front of oven with pans and bakeware .
A baker is working in the kitchen rolling dough .
A person standing by a stove in a kitchen .
A table with pies being made and a person standing near a wall with pots and pans hanging on the wall .

GroundTruth: dining table, oven, person, bottle, bowl, broccoli, carrot, cup, knife, sink, spoon
ExactMatch: dining table, oven, person
ExtendVocab: dining table, oven, person, pizza
GloVePseudo: oven
LearnedGloVe: dining table
TextClsf: person, oven, bowl, dining table, bottle, cup, spoon, knife, chair, refrigerator, pizza

A presenter projected on a large screen at a conference
People watching an on screen presentation of a gentleman in a suit .
People watch a man delivering a lecture on a screen .
A large screen showing a person wearing a suit
An audience is looking at an film of a man taking that is projected onto a wall .

GroundTruth: person, tie, bottle
ExactMatch: person
ExtendVocab: person
GloVePseudo: person
LearnedGloVe: person
TextClsf: person, tie, chair, handbag, tv

Fig. 6. Demonstration of different pseudo labels. Our method fills the gap between what is present and what is mentioned, by making inferences on the semantic level. Matches to the ground truth are shown in blue.

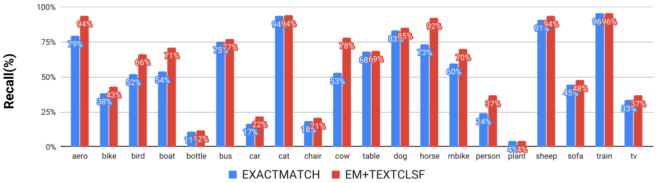


Fig. 7. Recall of PASCAL labels. We evaluate the recall of the COCO-learned text classifier, but we show only the overlapped 20 PASCAL VOC classes.

TABLE 2
COCO Test-Dev Results (Learning From COCO Captions), Measured by COCO Eval Server

Methods	Avg. Precision, IoU			Avg. Precision, Area			Avg. Recall, #Dets			Avg. Recall, Area		
	0.5:0.95	0.5	0.75	S	M	L	1	10	100	S	M	L
GT-Label	10.6	23.4	8.7	3.2	12.1	18.1	13.6	20.9	21.4	4.5	23.1	39.3
ExactMatch (EM)	8.9	19.7	7.1	2.3	10.1	16.3	**12.6**	**19.3**	**19.8**	3.4	20.3	37.4
EM + GloVe	8.6	19.0	6.9	2.2	10.0	16.0	12.2	18.7	18.9	2.9	19.0	37.6
EM + LearnedGloVe	8.9	19.7	7.2	2.5	10.4	**16.6**	12.3	19.1	19.6	**3.5**	20.0	37.7
EM + ExtendVocab	8.8	19.4	7.1	2.3	10.5	16.1	12.1	19.0	19.5	3.4	20.3	37.5
EM + TextClsf	**9.1**	**20.2**	**7.3**	**2.6**	**10.8**	**16.6**	12.5	**19.3**	**19.8**	**3.5**	**20.6**	**37.8**

Best method in bold. *Our EM+TextClsf achieves 86% of the GT-Label performance, and improves upon ExactMatch (EM) by 2.5%.*

For the results on COCO (Table 2), the gaps in performance between the different methods are smaller, but as before, our proposed EM+TextClsf shows the best performance. We believe the smaller gaps are because many of the COCO objects are not described precisely via natural language, and the dataset itself is more challenging than PASCAL thus gain may be diluted by tough examples.

Training With Flickr30K Captions. We also train our model on the Flickr30K [62] dataset, which contains 31,783 images and 158,915 descriptive captions. Training on Flickr30K is more challenging: on one hand, it includes less data compared to COCO; on the other hand, we observe that the recall rate of the captions is only 48.9% with ExactMatch which means only half of the data can be matched to some class names. In Table 1 (D), we observe that due to the limited training size, the detection models trained on Flickr30K captions achieve weaker performance than those trained on COCO captions. However, given the "free" supervision, the 33.6% mAP is still very encouraging. Importantly, we observe that even though our text classifier is trained on COCO captions and labels, it generalizes well to Flickr30K captions, as evidenced by the gap between EM+TextClsf and EM+ExtendVocab.

Results Without Pretraining on ImageNet. For our experiments thus far, we pretrain our visual backbone on ImageNet. This is a realistic setting consistent with WSOD because while ImageNet contains clean labels at the image level, no labels are available at the box level. However, to reduce the potential interference from those labels, we also test in a setting where no image-level labels are used to learn the visual representations. The results are shown in Table 1 (C, E). We observe that the advantage of our EM +TextClsf method remains. In particular, EM+TextClsf achieves 96% (=19.30/20.15) of the GT-Label COCO performance, which is comparable to (even higher than) the 93% achieved when pretraining on ImageNet was used (Table 1 A, B). Further, EM+TextClsf achieves a 5% gain (=19.30/18.46-1) over ExactMatch in Table 1 (C), versus 8% in (B). On Flickr30K, the gain of EM+TextClsf over ExactMatch is 12% without pretraining (E), greater than 8% with pretraining (D).

Training on Noisier Data: MIRFlickr1M and Conceptual Captions. The ultimate goal of our work is to enable training of object detection models from widely-available language data, e.g., user-generated content or narratives. The results thus far meet some of the challenges of working with language dataset as supervision, but rely on relatively clean datasets. Thus, we next extend our evaluation to two noisier datasets, MIRFlickr1M and Conceptual Captions. Examples from these datasets are shown in Figs. 4 and 11. Due to the noise in these datasets (see Figs. 4 and 11), performance is significantly lower when training on captions from them, compared to the cleaner COCO and Flickr30K. However, we observe that our EM+TextClsf still significantly improves upon the naive ExactMatch for both datasets, in Table 1 (F-H). In particular, EM+TextClsf *improves upon* ExactMatch *by 78% on MIRFlickr1M (subset), and 27-50% on Conceptual Captions (without/with ImageNet pretraining, G-H).* These are both much higher than the 8% gain of EM +TextClsf over ExactMatch, in segment Table 1 (B). We note the difference in performance when using a clean (COCO) versus noisy dataset (ConcCap). EM+TextClsf's performance on the latter is reduced by 37% compared to the former (27.2 mAP in G versus 43.1 in B), using the same amount of captions.

Our text classifier inferred reasonable image-level labels on noisier captions, even though it was trained on much cleaner data (COCO). For example, on Conceptual Captions, it was able infer 'person', 'cup', 'chair' and 'dining table' from the caption "front view of a young couple of college students drinking coffee and studying in a cafe", and 'person', 'tie', 'bottle', 'wine glass' and 'cup' from the caption "group of business people raising a toast with champagne at office". On MIRFlickr1M, it inferred 'tie' from the caption "For G. aaron kilt wedding whitby dunsleyhall", and 'knife', 'bowl', 'broccoli', 'carrot' from the caption "Roasted Veggies! cauliflower asparagus limes roasted food cooking kitchen vegetables dinner suppertime eating healthy explore". Thus, EM +TextClsf can infer labels that are implied but not stated, and leverage them for training.

Contribution of Pseudo Label Inference Using Alternative WSOD Method. For all of our experiments thus far, we have used OICR [8] as our WSOD method. In other words, once pseudo labels for the training set are inferred at the image level by one of the method alternatives, these labels are used to train a WSOD model, using OICR. Here, we experiment with replacing OICR with a more recent WSOD technique, namely Ren *et al.* [9]. We compare the performance of two label inference techniques, ExactMatch and EM +TextClsf, against using ground-truth labels at the image level (GT-Label). We show the results in Table 3, where all methods rely on Ren *et al.*'s WSOD method. We see that our proposed EM+TextClsf achieves 96% of the GT performance (using IoU 0.5), and a 4% boost over ExactMatch.

TABLE 3
Evaluation of Our Pseudo Label Inference, Using Ren *et al.* [9] as Our WSOD Method, on COCO2017-Val

Methods	Avg. Precision, IoU	
	0.5:0.95	0.5
GT-LABEL COCO	11.2	22.8
EXACT MATCH (EM)	10.0	21.1
EM + TEXTCLSF	10.4	22.0

These results are even better than Table 2 where our method achieved 86% of the ground-truth performance, and 2.5% gain over EXACTMATCH.

Data versus Performance. We show the potential of our model using Flickr30K and MIRFlickr1M [63]. For the latter, we concatenate the title and all user-generated content tags to form caption annotation. We then use our text classifier learned on COCO to rule out examples unlikely to mention our target classes. This filtering results in a dataset with around 20% of the original data, and we refer to it as Flickr200K. We use 10%, 20%, 50%, 100% data from both datasets, and report average precision on VOC 2007. We see from Fig. 9 that as training data increases, mAP increases accordingly. To estimate model potential, we fit a square root function to the rightmost four points in the figure and use it to estimate 54.4 mAP at 1 million samples.

Qualitative Results on COCO. We provide qualitative examples on the COCO *val* set, in Figs. 6 and 8. We compare EXACTMATCH and our EM+TEXTCLSF side-by-side in Fig. 8. Our proposed method EM+TEXTCLSF provides better detection results than the baseline EXACTMATCH. Thus, we conclude that it has squeezed more useful and precise information than the EXACTMATCH baseline.

4.4 Impact of Filtering Noisy Captions

Not all nouns in captions are object words, and not all object words are mentioned in the caption: for example, an image with caption “guests are sitting at a table during a wedding” will show object “table”, no object “wedding”, and additional objects (e.g., plates). The purpose of Section 3.1 is to determine which image-caption pairs contain significant alignment that would allow us to extract quality pseudo training labels at the image level. As discussed in Section 1, this is an alternative to our EM+TEXTCLSF (Section 3.2). The motivation for this alternative is that it does not require labels for training the text classifier. Thus, *our goal is for homogeneity scoring to improve our detection results with* EXACTMATCH, EM+EXTENDVOCAB *or* EM +GLOVE, not *with* EM+TEXTCLSF. Results with weighting on top of EM+TEXTCLSF are still shown, but marked with gray shading, in Tables 5, 6, 7, and 8.

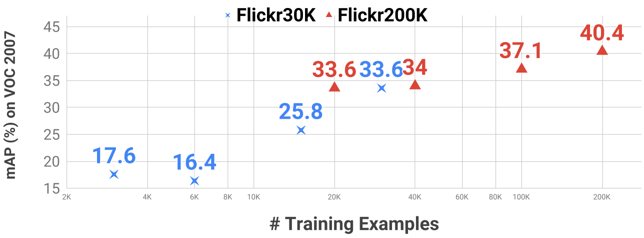


Fig. 9. Data versus Performance. Our text classifier learned on COCO generalized well on Flickr30K and the noisier Flickr200K data (subset of MIRFlickr1M) formed by user-generated content tags.

Ranking Captions by Potential Purity of Objects Mentions. We test to what extent homogeneity can be used to estimate if a caption contains clean labels. We use both captions and labels in COCO to compute a ground-truth *ranking of images* by the overlap between caption and object label words. We use the EXACTMATCH and EXTENDEDMATCH methods to compute the overlap, and rank images by precision (fraction of caption nouns that are also object labels). We then use our method to also compute an approximate ranking of images by the α^{HOM} scores, calculated individually for image and paired captions (Eq. (3)), as well as their difference. Finally, we compute the correlation between ground-truth rankings and our methods’ rankings, using Kendall’s τ and Spearman’s ρ. If the correlation is high, our method is a good indicator of how well-aligned a caption and object labels are, thus how likely it is that weakly-supervised detection will succeed if we extract labels from this particular caption.

We also compute approximate ranks using two baselines. HESSEL [40] computes visual concreteness for a word using the purity of images co-occurring with this word. We rank images by the average concreteness of nouns in their paired captions. ALIKHANI [39] collects annotations for the type of relation between an image and its caption, including “visible” (the most direct relation) and other less direct ones (e.g., “story”). We train a classifier using image and captions and the annotations from [39]. It outputs the probability of a caption being visible, and we sort images by their captions’ average visibility.

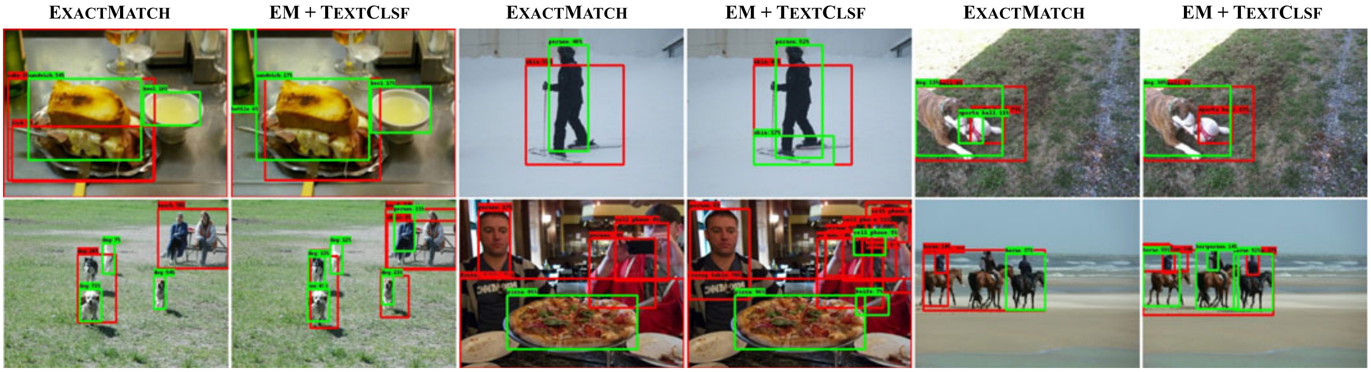


Fig. 8. Visualization of our Cap2Det model results on COCO *val* set. We show boxes with confidence scores $>$5%. Green boxes denote correct detection results ($IoU > 0.5$) while red boxes indicate incorrect ones. Best viewed with 300% zoom-in.

TABLE 4
Ranking Images and Object Categories by how Well Captions Overlap With the True Image-Level Labels

GT Ranking	Pred Ranking	Image		Label	
		ρ	τ	ρ	τ
ExactMatch	Hessel	**0.100**	**0.067**	0.293	0.182
	Alikhani	-0.081	-0.054	0.348	0.229
	Hom-Image	0.088	0.058	**0.382**	**0.230**
	Hom-Text	0.058	0.038	0.326	0.212
	Hom-Diff	-0.003	-0.003	0.003	-0.001
ExtendMatch	Hessel	0.043	0.029	0.229	0.142
	Alikhani	-0.009	-0.007	0.376	**0.251**
	Hom-Image	0.130	0.086	**0.388**	0.225
	Hom-Text	**0.179**	**0.120**	0.341	0.223
	Hom-Diff	0.177	0.118	0.053	0.037

Higher correlations are better. The best number per GT ranking is bolded. *Hom-Image is the best performer overall.*

TABLE 5
Comparing the Filtering Strategies With the Random Sampling Baseline, Using AP (in %) on VOC 2007 Test

Label inference Im-cap scoring	ExtendVocab	GloVe	TextClsf
Random (30K)	36.7	38.6	40.4
Hom-Image (30K)	40.1	**41.3**	**40.5**
Hom-Text (30K)	**40.4**	40.8	39.9

Both filtering mechanisms improve results under for ExtendVocab and GloVe pseudo label inference. Gray cells indicate we do not intend or expect improvement compared to no filtering.

TABLE 6
Evaluating the Caption Weighting Strategy, Using AP (in %) on VOC 2007 Test

Label inference Im-cap scoring	ExtendVocab	GloVe	TextClsf
No weighting (118K)	42.5	40.5	**43.1**
Hom-Image Weighting (118K)	**43.5**	**42.6**	42.2

We also *rank object categories*: by their concreteness scores for Hessel, by average visibility of images containing this category for Alikhani, and by average α^{HOM} values of images containing this category for our methods. In ground-truth rankings ExactMatch and ExtendedMatch, accuracy of category prediction (i.e., fraction of images where category c is correctly predicted) is used.

We show the results in Table 4. Hom-Image is the best performer overall, outperforming the alternatives in 3 cases (versus 2 for Hom-Text, 2 for Hessel, and 1 for Alikhani). We observe that the image ranking obtained by Hessel is more correlated with the ground-truth ranking acquired by ExactMatch. However, it fails to sustain its performance when ground-truth image ranking is ExtendMatch, probably due to not being able to capture visual concreteness for synonyms of object categories. In this setting, Hom-Text is the best. In the label ranking task, Hom-Image is the most correlated ranking overall.

Detection Results Using Image-Caption Filtering. We use a limited 30,000 image-caption subset from COCO *train2017* split for training, assuming a setting of restricted computation resources and training time. We keep the most useful examples while removing the others. We use the metric of *homogeneity* (Section 3.1.2) to measure the image-caption relevance. The higher this metric, the better alignment between the both modalities, and more likely the captions describe the visual objects in detail. Hom-Image and Hom-Text use $\alpha_{(I)}^{HOM}$ and $\alpha_{(T)}^{HOM}$ to filter examples, respectively. We use random sampling of 30K examples as a baseline.

Table 5 shows the results. We see the performances of ExtendVocab and GloVe are improved using the filtered training data. If we use a random selection of 30K examples, the performances are 36.7% and 38.6% respectively. Using *image homogeneity score* (Hom-Image, $\alpha_{(I)}^{HOM}$) for filtering improved these methods by 9% (40.1% versus 36.7%) and 7% (41.3% versus 38.6%). The TextClsf column is shaded in gray because we do not expect a boost in performance from filtering. TextClsf had already explained the gap between the image and text thus is not sensitive to the improved filtered training data. However, this text classifier requires a small number of ground-truth labels. In contrast, Hom-Image and Hom-Text with GloVe achieve competitive results to TextClsf with Random, but do not require any labels. Thus, homogeneity can be used to determine which captions provide strong supervision for object detection, without the need for *any* ground-truth labels. We omit Hessel, Alikhani and Hom-Diff in Table 5 because their overall performance in Table 4 does not exceed Hom-Image, while Alikhani also requires image-caption relation labels.

Results Using Image-Caption Weighting. One weakness of the filtering approach is that it requires a hard cutoff of the dataset examples. In comparison, weighting applies a soft "cutoff" to the data. It never drops data, thus is data-efficient. We use $\alpha_{(I)}^{HOM}$ as the per-example weighting factor γ to weigh the pseudo image-level labels extracted from different image-caption pairs in Eq. (9).

Comparing Tables 6 to 5, we see that even with a good filtering strategy, e.g., Hom-Image Filtering (30K), using 30K "clean" training examples is still inferior than training on the full COCO dataset (No weighting 118K)), for two of the three label inference methods (columns). However, if we apply the Hom-Image *weighting* on the loss, the performance (Hom-Image Weighting 118K)) is improved (43.5% versus 42.5%, 42.6% versus 40.5%, i.e., 2-5% gain), except for TextClsf which requires annotations. *This is an important finding with important ramifications for multimodal learning.* Approaches to learn visual representations have benefited greatly from widely available web/video data, and *our method suggests how the useful signal and noise in such data can be distinguished to boost the quality of the learned representations*, without requiring annotations. We note that ExtendVocab with weighting is comparable (only 1% better) to TextClsf with no weighting.

Results Using Large Corpus Pretraining. We conduct a study on the impact of vision-language pretraining, which has gained popularity for both VQA and cross-modal retrieval [25], [26]. However, no previous study had shown where such pre-trained models can also improve fully- or weakly-supervised object detection (except Zareian *et al.* [65] which use bounding boxes for some object categories). Our pre-training pipeline is shown in Fig. 10. We evenly distribute anchor boxes (different in scales and locations) across the image and treat each anchored visual region as a visual token. We then concatenate the visual tokens to the

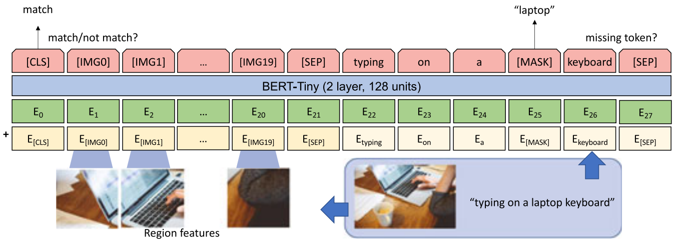


Fig. 10. Vision-language pretraining for weakly-supervised object detection. The inputs to the BERT-Tiny model are positional embeddings ($E_0, E_1, \ldots$) and token embeddings (proposal features ($E_{[IMG0]}$, $E_{[IMG1]}, \ldots$) and word embeddings ($E_{typing}, E_{on}, \ldots$)).

caption tokens, forming a sequence: "[CLS] [v1] [v2] ... [SEP] [t1] [t2] ... [SEP]", where the "[v]" and "[t]" are the visual and textual tokens, respectively. "[SEP]" is a special token to separate sequences, and "[CLS]" is the special classification token. We use the FastRCNN features and word embeddings to represent visual and text token embeddings, respectively. Then, we contextualize the sequence features using two self-attention layers (each with 128 hidden units). Finally, we add a linear classification layer on top of the "[CLS]" representation to predict a 0/1 value denoting if the image implies the caption's semantic meaning (matching). Besides the image-caption matching modeling, we also process the masked language modeling optimization. We randomly (with a probability of 15%) replace a text token with "[MASK]" and require the model to reconstruct the token, given the visual and text contexts. The visual model is the same as Section 3.4, while we use BERT-Tiny to initialize the text token embeddings and the self-attention layers. We use the Adam optimizer [53] with a batch size of 5, a learning rate of 1e-5, a weight decay of 1e-8, and we trim the maximum gradient norm to 1.0. Based on the above pretraining setting, we trained for 600K steps, roughly 1 epoch on the Conceptual Captions dataset. The training costs around 44 hours on 5 GeForce GTX1080Ti GPUs, using Tensorflow distributed training [57]. After pretraining the model using Conceptual Captions, we process the weakly supervised object detection training using the COCO images and texts and evaluate on the VOC07.

Table 7 shows the results. We see that pretraining improved the baseline by 3% (43.9% versus 42.5%), 2% (41.5% versus 40.5%), 1% (43.6% versus 43.1%). In the only pseudo label inference setting that requires neither a hand-defined vocabulary of synonyms, nor object labels, namely GLOVE, *our weighting achieves stronger results than pretraining*. For EXTENDVOCAB, pretraining and weighting achieve comparable results. Pretraining can be seen as a state-of-the-art method akin to Zareian *et al.* [65], as a way to use vision-language data, which could be alternative to our EM+TEXTCLSF. However, pretraining uses the external and large 3.3M Conceptual Captions dataset, while our EM+TEXTCLSF only uses the 118K COCO captions. Given inconclusive gains from pretraining over weighting (sometimes worse, sometimes better), and its large cost, it is not warranted in our setting. Applying both pretraining and weighting did not further boost results.

TABLE 7
The Impact of Large Corpus Pretraining, Measuring AP (in %) on VOC 2007 Test

Label inference Im-cap scoring	EXTENDVOCAB	GLOVE	TEXTCLSF
No weighting, no pretraining	42.5	40.5	43.1
+Weighting	43.5	**42.6**	42.2
+Pretraining	**43.9**	41.5	**43.6**
+Pretraining, +weighting	42.5	42.0	42.6

Both pretraining and weighting help, but weighting uses no external data.

TABLE 8
Weighting on Conceptual Captions, Using AP (in %) on VOC 2007 Test Set

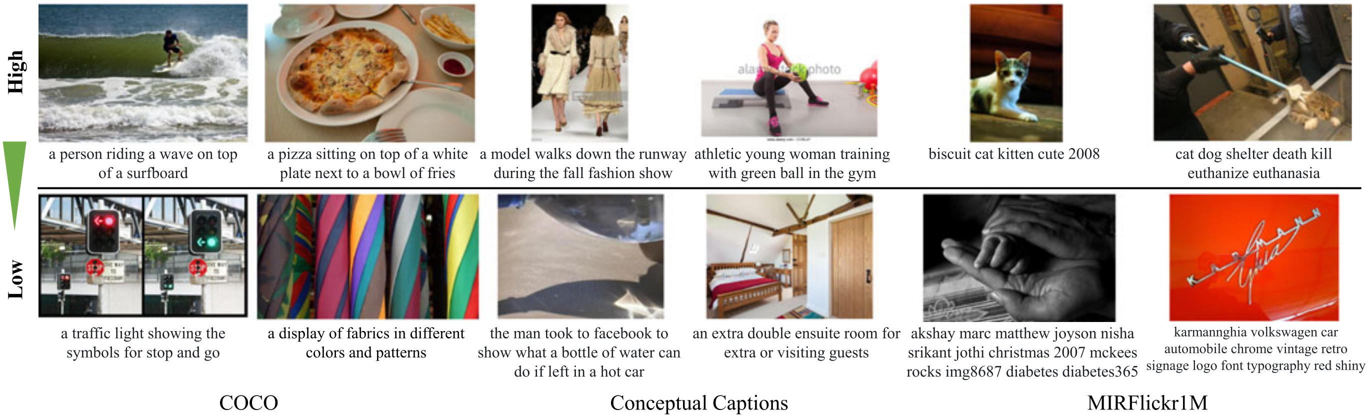
Label inference Im-cap scoring	EXACTMATCH	TEXTCLSF
(A) ImageNet pretraining in detection:		
No weighting	18.1	27.2
HOM-IMAGE Weighting (Triplet w/ IN PT)	24.7	18.8
HOM-IMAGE Weighting (Triplet)	24.0	25.5
(B) No ImageNet pretraining in detection:		
No weighting	2.38	3.02
HOM-IMAGE Weighting (Triplet)	2.99	2.75

Triplet models are trained on the COCO dataset. Weighting improves results across dataset boundaries. Gray cells indicate we do not intend or expect improvement compared to no weighting.

Generalization of Weighting on Conceptual Captions. The results described thus far in this subsection all apply filtering or weighting over COCO captions. The weighting model was trained on COCO, so in this part, we test the generalization of the weighting model, by applying it on the Conceptual Captions subset described in Section 4.1. Note that when applying weights to the captions, the gradient magnitude is reduced. Thus, rather than retuning the learning rate, we apply scaling to the weights such that the sum of weights over all samples remains the same as in the unweighted version. We include the result in Table 8. Part (A) uses ImageNet pretraining in extracting visual features for detection (Section 3.3), while part (B) does not. The second and third rows in (A) differ in the use of an ImageNet-trained backbone for the image-text model (trained with triplet loss). As before, our focus is on weighting improving the performance of the EXACTMATCH method. We observe that our weighting model does generalize to Conceptual Captions. On EXACTMATCH, we achieve a 33-37% boost in performance when using weighting (compared to no weighting) in (A), and 26% in (B). Thus, on Conceptual Captions, the impact of weighting is even more significant than in Table 6. TEXTCLSF is 10% better than EXACTMATCH with weighting (=27.2/24.7-1) but requires some labels.

Simplified Computation of Weights. The results in Table 8 replace the use of PVSE [21] to compute an image-text joint embedding, with a much simpler model trained with triplet loss and no attention. Thus, the gains from weighting that we achieve are not due to the complexity of PVSE.

Image-Caption Pairs With High/Low Scores. In Fig. 11, we show examples of image-caption pairs from each of COCO, Conceptual Captions, and MIRFlickr1M, that achieved high or low homogeneity scores. Among COCO samples, images with low homogeneity scores are usually more complex than ones with high scores, and mentioned concepts may be abstract ("symbols", "stop and go", "display", "fabrics"). On Conceptual Captions, high-scoring pairs describe the content



a person riding a wave on top of a surfboard

a pizza sitting on top of a white plate next to a bowl of fries

a model walks down the runway during the fall fashion show

athletic young woman training with green ball in the gym

biscuit cat kitten cute 2008

cat dog shelter death kill euthanize euthanasia

a traffic light showing the symbols for stop and go

a display of fabrics in different colors and patterns

the man took to facebook to show what a bottle of water can do if left in a hot car

an extra double ensuite room for extra or visiting guests

akshay marc matthew joyson nisha srikant jothi christmas 2007 mckees rocks img8687 diabetes diabetes365

karmannghia volkswagen car automobile chrome vintage retro signage logo font typography red shiny

COCO

Conceptual Captions

MIRFlickr1M

Fig. 11. Image-caption pairs with high homogeneity scores on the top, and low scores on the bottom.

TABLE 9
Average Precision (in %) on the Pascal VOC Test Set Using *Ground-Truth* Image-Level Labels

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
VOC 2007 results:																					
OICR VGG16 [8]	58.0	62.4	31.1	19.4	13.0	**65.1**	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	**24.1**	41.7	46.9	**64.3**	62.6	41.2
PCL-OB-G VGG16 [16]	54.4	**69.0**	39.3	19.2	**15.7**	62.9	**64.4**	30.0	**25.1**	52.5	44.4	19.6	39.3	**67.7**	**17.8**	22.9	46.6	**57.5**	58.6	63.0	43.5
TS^{2}C [7]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
OICR Ens.+FRCNN [8]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	*63.1*	49.5	30.3	*64.7*	66.1	13.0	*25.6*	*50.0*	57.1	60.2	59.0	47.0
PCL-OB-G Ens.+FRCNN [16]	63.2	*69.9*	47.9	22.6	*27.3*	*71.0*	*69.1*	49.6	12.0	60.1	*51.5*	37.3	63.3	63.9	15.8	23.6	48.8	55.3	61.2	62.1	*48.8*
Ours	**68.7**	49.7	**53.3**	**27.6**	14.1	64.3	58.1	**76.0**	23.6	**59.8**	**50.7**	**57.4**	**48.1**	63.0	15.5	18.4	**49.7**	55.0	48.4	**67.8**	**48.5**
VOC 2012 results:																					
OICR VGG16 [8]	67.7	61.2	41.5	25.6	22.2	54.6	49.7	25.4	19.9	47.0	18.1	26.0	38.9	67.7	2.0	22.6	41.1	34.3	37.9	55.3	37.9
PCL-OB-G VGG16 [16]	58.2	**66.0**	41.8	24.8	**27.2**	55.7	**55.2**	28.5	16.6	**51.0**	17.5	28.6	**49.7**	**70.5**	7.1	**25.7**	47.5	36.6	44.1	**59.2**	40.6
TS^{2}C [7]	67.4	57.0	37.7	23.7	15.2	**56.9**	49.1	64.8	15.1	39.4	19.3	**48.4**	44.5	67.2	2.1	23.3	35.1	40.2	**46.6**	45.8	40.0
OICR Ens.+FRCNN [8]	71.4	69.4	55.1	29.8	*28.1*	55.0	*57.9*	24.4	17.2	*59.1*	21.8	26.6	57.8	71.3	1.0	23.1	*52.7*	37.5	33.5	56.6	42.5
PCL-OB-G Ens.+FRCNN [16]	69.0	*71.3*	*56.1*	30.3	27.3	55.2	57.6	30.1	8.6	56.6	18.4	43.9	*64.6*	*71.8*	7.5	23.0	46.0	44.1	42.6	58.8	44.2
Ours	**74.2**	49.8	**56.0**	**32.5**	22.0	55.1	49.8	**73.4**	**20.4**	47.8	**32.0**	39.7	48.0	62.6	**8.6**	23.7	**52.1**	**52.5**	42.9	59.1	**45.1**

The top shows VOC 2007 and the bottom shows VOC 2012 results. The best single model is in bold, *and best ensemble in* italics.

in a literal fashion, and many objects are mentioned. In low-scoring pairs, mentioned objects are not present (e.g., "man") or present objects are not mentioned (e.g., the bed). On MIRFlickr1M, the high-scoring images mention object categories, while the low-scoring ones are significantly more abstract or non-object-like. Thus, we qualitatively showed that the homogeneity scores measure the relevance and redundancy between the image and text modalities, in terms of ability to extract object labels. Homogeneity helps to rule out less useful examples to better train a detector.

4.5 Verifying WSOD With Ground-Truth Image Labels

We finally show the performance of our method in the classic WSOD setting where *clean image-level labels for training are available* and do not need to be inferred. These results validate the method component in Section 3.3. Our goal is not to exceed the very latest WSOD models, but to perform on par with recent ones. Note that all methods tested in this section, including ours, use *ground-truth image-level labels*, but differ in terms of architectures and WSOD techniques. We refer to our method here as a WSOD variant, to distinguish it from the main focus of our work on using language supervision (which is not utilized here). Also note that the multi-scale training and test time augmentation mentioned in our Section 3.4 are widely adopted in WSOD. We verified that all baseline methods in this section use them. Thus, our comparisons to the SOTA methods in this section are fair.

Results on PASCAL VOC. For each image, we extract object categories from all the ground-truth bounding boxes, and only keep these *image-level* labels for training, discarding box information. For VOC 2007 and 2012, we train on 5,011 and 11,540 *trainval* images respectively and evaluate on 4,952 and 10,991 *test* images. We report mean Average Precision (mAP) at IoU > 0.5, and compare against multiple strong WSOD baselines, in Table 9. The WSOD variant of our model performs on par with or better than the baselines on both VOC 2007 and 2012.

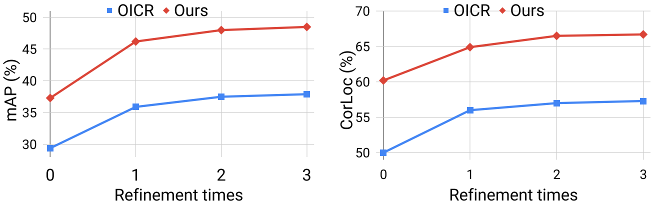


Fig. 12. Analysis of our basic network and OICR components on VOC 2007. Comparison of the performance of our model and OICR VGG_M after iterative refinement.

TABLE 10
COCO Detection Using *Ground-Truth* Image Labels, With Supervised Models at the Top, Best WSOD in *bold*

Methods	Avg. Precision, IoU	
	0.5:0.95	0.5
Faster RCNN [1]	21.9	42.7
Faster Inception-V2 [59]	28.0	-
PCL-OB-G VGG16 [16]	8.5	19.4
PCL-OB-G Ens.+FRCNN [16]	9.2	19.6
Ours	**10.6**	**23.4**

Effects of the Basic Network and OICR. The performance gain of our model comes from two aspects: (1) a more advanced detection model backbone architecture and (2) the online instance classifier refinement (OICR). Fig. 12 shows the performance of the WSOD variant of our method and that of Tang [8] (OICR VGG_M), both refining for 0, 1, 2, 3 times. With no (0) refinement, our basic network architecture outperforms the VGG_M backbone of Tang *et al.* by 27% in mAP. But the basic architecture improvement is not sufficient to achieve top results. If we use OICR to refine the models 1, 2, or 3 times, we gain 24%, 29%, and 30% respectively while Tang achieve smaller improvement (22%, 28%, and 29% gains).

Results on COCO. We train the WSOD variant of our model on the 118,287 *train2017* images, using the image-level ground truth labels. We report mAP at IoU=.50:.05:.95 and mAP@0.5, on the 20,288 *test-dev2017* images. We compare to a representative fully-supervised detection model [1]; "Faster Inception-V2" [58] which is our WSOD variant's supervised detection counterpart (using bounding-box annotations), and a recent WSOD model, PCL-OB-G Ens + FRCNN [16]. As demonstrated in Table 10, our model outperforms the WSOD model by 15% in terms of mAP, but as expected, the gap between with the supervised method is still large due to the disparate supervision strength.

5 Conclusion

We showed how we can successfully leverage naturally arising, weak supervision in the form of captions. We amplify the signal that captions provide by learning to bridge the gap between what human annotators mention, and what is present in the image. We also learn how to weigh the contribution of different captions as supervision, based on the expected alignment between the image and caption. In the future, we will extend our method to incorporate raw supervision in the form of spoken descriptions in video.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[3] K. Ye, A. Kovashka, M. Sandler, M. Zhu, A. Howard, and M. Fornoni, "SpotPatch: Parameter-efficient transfer learning for mobile object detection," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 636–640.

[4] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3339–3348.

[5] C. Thomas and A. Kovashka, "Artistic object recognition by unsupervised style adaptation," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 460–476.

[6] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2846–2854.

[7] Y. Wei *et al.*, "Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 434–450.

[8] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3059–3067.

[9] Z. Ren *et al.*, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 595–10 604.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[11] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-MIL: Continuation multiple instance learning for weakly supervised object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2194–2203.

[12] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8291–8299.

[13] Z. Ren *et al.*, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10595–10604.

[14] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 685–694.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[16] P. Tang *et al.*, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[17] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5131–5139.

[18] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1297–1306.

[19] S. K. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3270–3277.

[20] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 35–44.

[21] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1979–1988.

[22] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[23] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[24] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 451–466.

[25] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[26] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 5100–5111.

[27] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 817–834.

[28] D. Surís, D. Epstein, H. Ji, S.-F. Chang, and C. Vondrick, "Learning to learn words from visual scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 434–452.

[29] C. Thomas and A. Kovashka, "Preserving semantic neighborhoods for robust cross-modal retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 317–335.

[30] L. Gomez, Y. Patel, M. Rusiñol, D. Karatzas, and C. V. Jawahar, "Self-supervised learning of visual features through embedding images into text topic spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2017–2026.

[31] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9876–9886.

[32] G. Bertasius and L. Torresani, "COBE: Contextualized object embeddings from narrated instructional video," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 1269.

[33] K. Desai and J. Johnson, "VirTex: Learning visual representations from textual annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11 157–11 168.

[34] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[35] K. Chen, H. Song, C. C. Loy, and D. Lin, "Discover and learn new objects from documentaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1111–1120.

[36] K. Ye, M. Zhang, A. Kovashka, W. Li, D. Qin, and J. Berent, "Cap2Det: Learning to amplify weak caption supervision for object detection," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9685–9694.

[37] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2930–2939.

[38] M. Zhang, R. Hwa, and A. Kovashka, "Equal but not the same: Understanding the implicit relationship between persuasive images and text," in *Proc. Brit. Mach. Vis. Conf.*, 2018, Art. no. 8.

[39] M. Alikhani, P. Sharma, S. Li, R. Soricut, and M. Stone, "Clue: Cross-modal coherence modeling for caption generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6525–6535.

[40] J. Hessel, D. Mimno, and L. Lee, "Quantifying the visual concreteness of words and topics in multimodal datasets," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 2194–2205.

[41] C. Thomas and A. Kovashka, "Emphasizing complementary samples for non-literal cross-modal retrieval," in *Proc. Multimodal Learn. Appl. Workshop*, 2022.

[42] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[43] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[45] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[46] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7219–7228.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[48] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[50] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. 8th Workshop Syntax Semantics Struct. Statist. Transl.*, 2014, pp. 103–111.

[51] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC Workshop New Challenges NLP Frameworks*, 2010, pp. 45–50. [Online]. Available: http://is.muni.cz/publication/884893/en

[52] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[54] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, Apr. 2020.

[55] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.

[56] G. Bradski, "The openCV library," *Dr. Dobb's J. Softw. Tools*, vol. 25, pp. 120–123, 2000.

[57] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. USENIX Conf. Operating Syst. Des. Implementation*, 2016, pp. 265–283.

[58] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3296–3297.

[59] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[60] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[61] X. Chen *et al.*, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

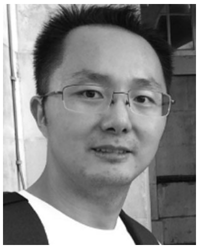
[62] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.

[63] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: The MIR flickr retrieval evaluation initiative," in *Proc. Int. Conf. Multimedia Inf. Retrieval*, 2010, pp. 527–536.

[64] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.

[65] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 393–14 402.

Mesut Erhan Unal received the BSc degree in computer science and engineering from Hacettepe University, Turkey. He is currently working toward the PhD degree in computer science with the University of Pittsburgh (Pitt). Before starting his PhD at Pitt, he worked for Jotform as a software engineer for two years. The goal of his research is to better understand, interpret and manipulate visual data using natural language supervision and contextual information.



Keren Ye received the bachelor's and master's degrees from Beihang University, China, in 2004-2011, and the PhD degree in computer science from the University of Pittsburgh, in August 2021. He is now a senior applied research scientist with Cruise, San Francisco. His research interests include object detection, multi-modal learning, and knowledge representation. Before studying at Pitt, he worked as a software engineer with Baidu Inc. for five years.

Mingda Zhang received the BSc degree in chemical biology from Peking University, China, in 2013, and the PhD degree in computer science from the University of Pittsburgh, in December 2021. He is currently a software engineer with Google Research, New York. His research interests include the intersection of computer vision and natural language processing. He has completed research internships in Google AI, Seattle.

Christopher Thomas received the PhD degree in computer science from the University of Pittsburgh, in August 2020. He is currently an assistant professor in computer science with Virginia Tech. Previously, he was a postdoctoral researcher with Columbia University, mentored by Prof. Shih-Fu Chang and working as part of the DARPA SemaFor program. His research interests include semantic and pragmatic image understanding, structured, weak, and unsupervised learning, vision and language, and image generation.

Adriana Kovashka received the PhD degree from the University of Texas at Austin, in August 2014. She is currently an associate professor in computer science with the University of Pittsburgh. Her work has been published in CVPR, ICCV, ECCV, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, NeurIPS, AAAI, and ACL, and has been funded by NSF, Google, Amazon, and Adobe. She is the recipient of a NSF CAREER Award in 2021. She served as area chair for CVPR 2018-2021, and will serve as ICCV 2025 program co-chair.

Wei Li received the bachelor's degree from Tsinghua University, and the master's degree from CUHK. He is a research scientist with NewsBreak Seattle. His current research focus is on video understanding, editing, and synthesis. His work has been published at CVPR, ICCV, *IEEE Transactions on Pattern Analysis and Machine Intelligence* and etc.

Danfeng Qin received the PhD degree from the Computer Vision Lab, ETH Zurich. She is currently a software engineer with Google AI. Her research interests include label efficient learning with web data.

Jesse Berent received the master's degree in microengineering from the EPFL, Switzerland, and the PhD degree in communications and signal processing from Imperial College London, in 2008. He is currently a research scientist and tech lead manager with Google AI Zurich. His group focuses on image and video analysis, digital ink recognition, and machine learning. Prior to joining Google in 2009, he was a postdoctoral researcher with Imperial College London focusing on multi-view image analysis.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.