# Localized Vision-Language Matching for Open-vocabulary Object Detection

María A. Bravo, Sudhanshu Mittal, and Thomas Brox

Department of Computer Science
University of Freiburg, Germany
{bravoma, mittal, brox}@cs.uni-freiburg.de

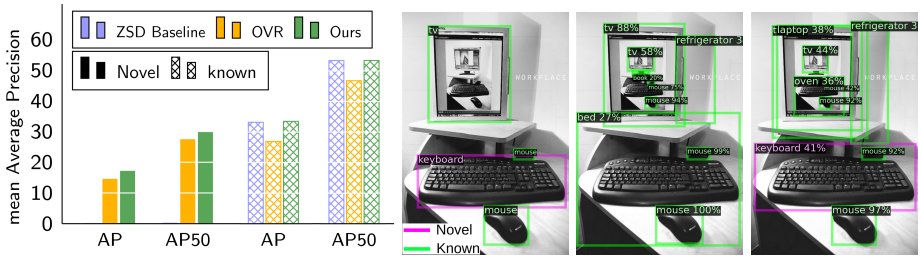
**Abstract.** In this work, we propose an open-vocabulary object detection method that, based on image-caption pairs, learns to detect novel object classes along with a given set of known classes. It is a two-stage training approach that first uses a location-guided image-caption matching technique to learn class labels for both novel and known classes in a weakly-supervised manner and second specializes the model for the object detection task using known class annotations. We show that a simple language model fits better than a large contextualized language model for detecting novel objects. Moreover, we introduce a consistency-regularization technique to better exploit image-caption pair information. Our method compares favorably to existing open-vocabulary detection approaches while being data-efficient. Source code is available at https://github.com/lmb-freiburg/locov.

**Keywords:** Open-vocabulary Object Detection, Image-caption Matching, Weakly-supervised Learning, Multi-modal Training

## 1 Introduction

Recent advances in deep learning have rapidly advanced the state-of-the-art object detection algorithms. The best mean average precision score on the popular COCO [23] benchmark has improved from 40 mAP to over 60 mAP in less than 4 years. However, this success required large datasets with annotations at the bounding box level and was a achieved in a closed-world setting, where the number of classes is assumed to be fixed. The closed-world setting restricts the object detector to only discover known annotated objects and annotating all possible objects in the world is infeasible due to high labeling costs. Therefore, research of open-world detectors, which can also discover unmarked objects, has recently come into focus specially using textual information together with images for open-vocabulary detection [13, 40, 43].

To learn a visual concept, humans receive the majority of the supervision in the form of narrations rather than class tags and bounding boxes. Consider the example of Figure 1 together with the annotations of mouse and tv only. Even after learning to detect these objects, finding and identifying the keyboard without any other source of information is ambitious. Instead, if we consider




(a) SOTA method comparison (b) GT (c) ZSD (d) LocOv

Fig. 1: Open-vocabulary object detection. (a) Compares our method LocOv with the baseline method (OVR) and our zero-shot baseline STT-ZSD (ZSD). LocOv improves on both novel and known classes without dropping the performance on known classes. The zero-shot method, only trained with known classes, obtains low performance ($< 0.5$ mAP) on novel classes. (b-d) LocOv is able to detect the novel object 'keyboard' along with known objects, shown in figure.

the image together with the caption - "A mouse, keyboard, and a monitor on a desk", it is possible to identify that the other salient object in the image is very likely a keyboard. This process involves successful localization of the objects in the scene, identification of different nouns in the narrated sentence, and matching the two together. Exploiting the extensive semantic knowledge contained in natural language is a reasonable step towards learning such open-vocabulary models without expensive annotation costs.

In this work, we aim to learn novel objects using image-caption pairs. Along with image-caption pairs, the detector is provided with box annotations for a limited set of classes. We follow the problem setting as introduced by Zareian *et al.* [40]. They refer to this problem as *Open-vocabulary Object Detection*. There are two major challenges to this problem: First, image-caption pairs themselves are too weak to learn localized object-regions. Analyzing previous works, we find that randomly sampled feature maps provide imprecise visual grounding for foreground objects, therefore they receive insufficient supervisory signals to learn object properties. Second, the granularity of the information captured by image-region features should align with the level of information captured by the text representation for an effective matching. For example, it would be ill-suited to match a text representation that captures global image information with image features that capture localized information.

In this work, we propose a method that improves the matching between image and text representations. Our model is a two-stage approach: in the first stage, *Localized Semantic Matching* (LSM), it learns semantics of objects in the image by matching image-regions to the words in the caption; and in the second stage, *Specialized Task Tuning* (STT), it learns specialized visual features for the target object detection task using known object annotations. We called our method LocOv for **Loc**alized Image-Caption Matching for **O**pen-**v**ocabulary.

For the given objects in an image, our goal is to project them to a feature space where they can be matched with their corresponding class in the form of text embeddings. We find that simple text embeddings are better candidates for matching object representations than contextualized embeddings produced by large-scale language models.

Using image-caption pairs as weak supervision for object detection requires the understanding of both modalities in a fine and a coarse way. This can be obtained by processing each modality independently in a uni-modal fashion and then matching, or using cross-modal attention to process them together. To ensure consistent training between the uni-modal and cross-modal methods, we propose a consistency-regularization between the two matching scores. To summarize, our contributions are: (1) We introduced localized-regions during the image-caption matching stage to improve visual feature learning of objects. (2) We show that simplified text embeddings match better with identified object features as compared to contextualized text embeddings. (3) We propose a consistency regularization technique to ensure effective cross-modal training.

These three contributions allow LocOv to be not only competitive against state-of-the-art models but also data-efficient by using less than 0.6 million image-caption pairs for training, $\sim$700 times smaller than CLIP-based methods. Additionally, we define an open-vocabulary object detection setup based on the VAW [28] dataset, which offers challenging learning conditions like few-instances per object and a long-tailed distribution. Based on the above mentioned three contributions, we show that our method achieves state-of-the-art performance on both open-vocabulary object detection benchmarks, COCO and VAW.

## 2 Related Work

**Object detection with limited supervision** Semi-supervised (SSOD) [17, 24, 34] and weakly-supervised (WSOD) [4, 8, 20] object detection are two widely explored approaches to reduce the annotation cost. WSOD approaches aim to learn object localization using image-level labels only. Major challenges in WSOD approaches include differentiation between object instances [32] and precisely locating the entire objects. SSOD approaches use a small fully-annotated set and a large set of unlabeled images. Best SSOD [24, 34] methods are based on pseudo-labeling, which usually suffers from foreground-background imbalance and overfitting on the labeled set of images. In this work, we address a problem which shares similar challenges with the WSOD and SSOD approaches, however they are limited to a closed-world setting with a fixed and predefined set of classes. Our method addresses a mixed semi- and weakly-supervised object detection problem where the objective is open-vocabulary object detection.

**Multi-modal visual and language models.** Over the past years, multiple works have centered their attention on the intersection of vision and language by exploiting their consistent semantic information contained in matching pairs. The success of using this pairwise information has proved to be useful for pre-training transformer-like models for various vision-language tasks [6, 21, 25, 35,

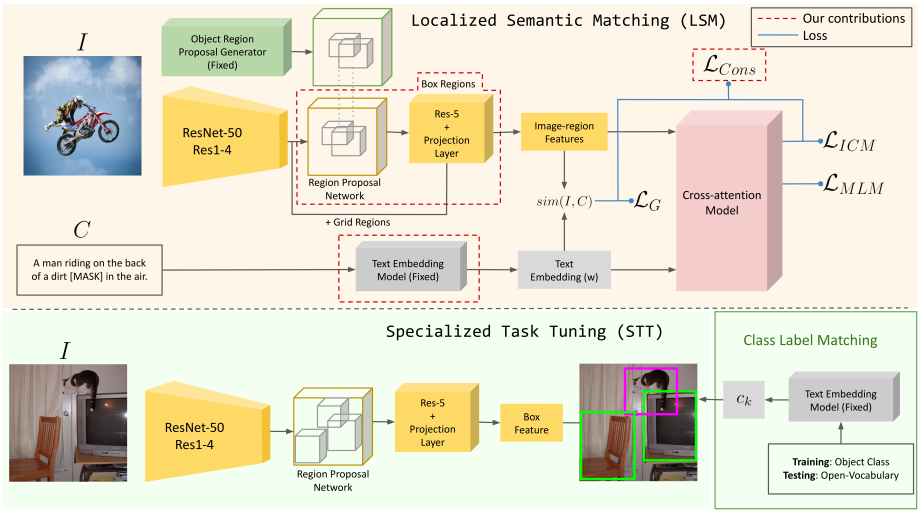



Fig. 2: Overview of LocOv . It is a two-stage model: (1) Localized Semantic Matching stage trains a Faster R-CNN-based model to match corresponding image-caption pairs using a grounding loss $\mathcal{L}_G$. We exploit the multi-modal information by using a cross-attention model and an Image-Caption matching loss $\mathcal{L}_{ICM}$, the mask language modeling loss $\mathcal{L}_{MLM}$ and a consistency-regularization loss $\mathcal{L}_{Cons}$. (2) Specialized Task Tuning stage tunes the model using the known class annotations and specializes the model for object detection. See Section 3.

36, 41, 44] which process the information jointly using cross-attention. Other approaches [11,12,19,26,29,37], centered on the vision and language retrieval task use separate encoders for each modality, in a uni-modal fashion. These models give the flexibility to transfer the knowledge learned by the pairwise information to single modality tasks, which is the case of object detection. In particular Miech *et al.* [26] showed that combining a cross-attention model with two uni-modal encoders is beneficial for large-scale retrieval tasks. In this paper, we combine the strengths of both types of approaches to train a model using different consistency losses that exploit the information contained in image-caption pairs.

**Language-guided object detection.** Zero-shot object detection methods learn to align proposed object-region features to the class-text embeddings. Bansal *et al.* [2] is among the first to propose the zero-shot object detection problem. They identified that the main challenge in ZSD is to separate the background class from the novel objects. Zhu *et al.* [45] trained a generative model to "hallucinate" (synthesize visual features) unseen classes and used these generated features during training to be able to distinguish novel objects from background. Rahman *et al.* [30] proposed a polarity loss to handle foreground-background imbalance and to improve visual-semantic alignment. However, such methods fail to perform well on the novel classes since the detection model has

never seen these novel objects, and semantics learned by matching known object-text embeddings does not extrapolate to novel classes.

To learn the semantics of novel classes, recent methods [3, 13, 16, 40, 43] have simplified the problem by providing image-caption pairs as a weak supervision signal. Such pairs are cheap to acquire and make the problem tractable. Image-caption pairs allow the model to observe a large set of object categories along with object labels. These methods either use this model to align image-regions with captions and generate object-box pseudo labels [16, 43] or as region-image feature extractor to classify the regions [13]. Many weakly-supervised [1, 3, 7, 33, 42] approaches have been proposed to perform such object grounding. Due to the large performance gap between zero-shot/weakly-supervised and fully-supervised approaches for object detection, Zareian *et al.* [40] introduced an open-vocabulary problem formulation. It utilizes extra image-caption pairs to learn to detect both known and novel objects. Their approach matches all parts of the image with the caption, whereas we emphasize object localized regions and a consistency loss to enforce more object-centric matching.

## 3 Method

We propose a two-stage approach for the task of open-vocabulary object detection as shown in Figure 2. The first stage, *Localized Semantic Matching* (LSM), learns to match objects in the image to their corresponding class labels in the caption in a weakly-supervised manner. The second stage, *Specialized Task Tuning* (STT) stage, includes specialized training for the downstream task of object detection. We consider two sets of object classes: known classes $O_K$ and novel classes $O_N$. Bounding box annotations, including class labels, are available for known classes whereas there are no annotations for the novel classes.

The LSM receives image-caption pairs $(I, C)$ as input, where the caption provides the weak supervision to different image-regions. Captions contain rich information which often include words corresponding to object classes from both known and novel sets. Captions are processed using a pre-trained text-embedding model (*e.g.*BERT [10] embedding) to produce word or part-of-word features. Images are processed using an object detection network (Faster R-CNN [31]) to obtain object region features. We propose to utilize an object proposal generator OLN [18] to provide regions as pseudo-labels to train the Faster R-CNN. This helps obtaining object-rich regions which improve image region-caption matching. This way, during the LSM our model learns to match all present objects in the image in a class-agnostic way. See Section 3.1 for details. The STT stage tunes the Faster R-CNN using known object annotations primarily to distinguish foreground from background and learns corresponding precise location of the foreground objects. See Section 3.2 for details.

### 3.1 Localized Semantic Matching (LSM)

The LSM stage consists of three main components: (1) localized object region-text matching, (2) disentangled text features and (3) consistency-regularization.

**Localized object region-text matching.** Given the sets $R^I = \{\ r : r$ is an image-region feature vector from the image I$\}$ and $W^C = \{\ w : w$ is a word or part-of-word feature vector from the caption C$\}$, we calculate the similarity score between an image and a caption in a fine-grained manner, by comparing image-regions with words, since our final objective is to recognize objects in regions. The image is processed using a Faster R-CNN model and a projection layer that maps image-regions into the text-embedding feature space. The similarity score is calculated by taking an image composed of $|R^I|$ region features and a caption composed of $|W^C|$ part-of-word features by:

$$sim(I, C) = \frac{1}{|R^I|} \sum_{i=1}^{|R^I|} \sum_{j=1}^{|W^C|} d_{i,j}(r_i \cdot w_j) \tag{1}$$

where $d_{i,j}$ corresponds to:

$$d(r_i, w_j) = d_{i,j} = \frac{\exp(r_i \cdot w_j)}{\sum_{j'=1}^{|W^C|} \exp(r_i \cdot w_{j'})}. \tag{2}$$

Based on the similarity score (Eq. 1) , we apply a contrastive learning objective to match the corresponding pairs together by considering all other pairs in the batch as negative pairs. We define this grounding loss as:

$$\mathcal{L}_{G_r}(I) = -\log \frac{\exp(sim(I, C))}{\sum_{C' \in \text{Batch}} \exp(sim(I, C'))} \tag{3}$$

We apply this loss in a symmetrical way, where each image in the batch is compared to all captions in the batch (Eq. 3) and each caption is compared to all images in the batch $\mathcal{L}_{G_r}(C)$. The subscript $r$ denotes the type of image-regions used for the loss calculation. We consider two types of image-regions: box-regions and grid-regions. Box-region features are obtained naturally using the region of interest pooling (RPN) from the Faster R-CNN. We make use of the pre-trained object proposal generator (OLN) to train the Faster-RCNN network. OLN is a class-agnostic object proposal generator which estimates all objects in the image with a high average recall rate. We train OLN using the known class annotations and use the predicted boxes to train our detection model, shown in Figure 2. Since captions sometimes refer to background context in the image, parallel to the box-region features, we also use grid-region features similar to the OVR [40] approach. Grid-region features are obtained by skipping the RPN in the Faster R-CNN and simply using the output of the backbone network. We apply the grounding loss to both type of image-region features. Our final grounding loss is given by:

$$\mathcal{L}_G = \mathcal{L}_{G_{box}}(C) + \mathcal{L}_{G_{box}}(I) + \mathcal{L}_{G_{grid}}(C) + \mathcal{L}_{G_{grid}}(I) \tag{4}$$

**Disentangled text features.** Many previous works [6, 15, 25, 35] use contextualized language models to extract text representations of the sentence. Although, this might be suitable for a task that requires a global representation of

a phrase or text, this is not ideal for the case for object detection, where each predicted bounding box is expected to contain a single object instance. We show that using a simple text representation, which keeps the disentangled semantics of words in a caption, gives the flexibility to correctly match object boxes in an image with words in a caption. Our method uses only the embedding module [10, 27] of a pre-trained language model to encode the caption and perform matching with the proposed image-regions. For embedding model we refer to the learned dictionary of vector representations of text tokens, which correspond to words or part-of-words. For cases where the text representing an object category is divided into multiple tokens, we consider the average representation of the tokens as the global representation of the object category. We show empirically, in Section 4.4, that using such a lightweight text embedding module has better performance than using a whole large-scale language model.

**Consistency-regularization** Miech *et al.* [26] showed that processing multi-modal data using cross-attention networks brings improvements in retrieval accuracy over using separate encoders for each modality and projecting over a common embedding space. However, this cross-attention becomes very expensive when the task requires large-scale retrieval. To take the benefit of cross-attention models, we consider a model similar to PixelBERT [15] to process the image-caption pairs. This cross-attention model takes the image-regions $R^I$ together with the text embeddings $W^C$ and matches the corresponding image-caption pairs in a batch. The image-caption matching loss ($\mathcal{L}_{ICM}$) of the cross-attention model together with the traditional Masking Language Modeling loss ($\mathcal{L}_{MLM}$) enforces the model to better project the image-region features to the language semantic space.To better utilize the cross-attention model, we propose a consistency-regularization loss ($\mathcal{L}_{Cons}$) between the final predicted distribution over the image-caption matching scores in the batch, before and after the cross-attention model. We use the Kullback-Leibler divergence loss to impose this consistency. In summary, we use three consistency terms over different image-caption pairs:

$$\begin{aligned}\mathcal{L}_{Cons} =& D_{KL}(p(I_{box}, C)||q(I_{box}, C)) \\ &+ D_{KL}(p(I_{grid}, C)||q(I_{grid}, C)) \\ &+ D_{KL}(p(I_{grid}, C)||q(I_{box}, C))\end{aligned} \tag{5}$$

where $p(I_*, C)$ and $q(I_*, C)$ correspond to the softmax of the image-caption pairs in a batch before and after the cross-attention model respectively, and the sub-index of the image corresponds to the box- or grid-region features. Our final loss for the LSM stage corresponds to the sum of the above defined losses:

$$\mathbf{L}_{LSM} = \mathcal{L}_G + \mathcal{L}_{ICM} + \mathcal{L}_{MLM} + \mathcal{L}_{Cons} \tag{6}$$

### 3.2 Specialized Task Tuning (STT)

In this stage, we fine-tune the model using known class annotations to learn to localize the objects precisely. We initialize the weights from the LSM stage model,

and partially freeze part of the backbone and the projection layer to preserve the learned semantics. Freezing the projection layer is important to avoid overfitting on the known classes and generalize on novel classes. To predict the class of an object, we compute the similarity score between the proposed object box-region feature vector ($r_i$) and all the class embedding vectors $c_k$ and apply softmax

$$p(r_i, c_k) = \frac{\exp(r_i \cdot c_k)}{1 + \sum_{c'_k \in O_K} \exp(r_i \cdot c_{k'})}. \tag{7}$$

The scalar 1 included in the denominator corresponds to the background class, which has a representation vector of all-zeros. We evaluate the performance across three setups: (Novel) considering only the novel class set $O_N$, (Known) comparing with the known classes only $O_K$ and (Generalized) considering all novel and known classes together.

## 4 Experiments

### 4.1 Training Details

**Datasets.** The **Common Objects in Context (COCO) dataset** [22] is a large-scale object detection benchmark widely used in the community. We use the 2017 train and val split for training and evaluation respectively. We use the known and novel object class splits proposed by Bansal *et al.* [2]. The known set consists of 48 classes while the novel set has 17 classes selected from the total of 80 classes of the original COCO dataset. We remove the images which do not contain the known class instances from the training set. For the localized semantic matching phase, we use the captions from **COCO captions [5]** dataset which has the same train/test splits as the COCO object detection task. COCO captions dataset contains 118,287 images with 5 captions each. Additionally in the supplementary material, we test LocOv using **Visual Attributes in the Wild (VAW) dataset [28]** a more challenging dataset containing fine-grained classes with a long-tailed distribution.

**Evaluation metric.** We evaluate our method using mean Average Precision (AP) over IoU scores from 0.5 to 0.95 with a step size of 0.05, and using two fixed thresholds at 0.5 ($AP_{50}$) and 0.75 ($AP_{75}$). We compute these metrics separately for novel and known classes, calculating the softmax within the subsets exclusively; and in a generalized version both sets are evaluated in a combined manner, calculating the probability across all classes.

**Implementation details.** We base our model on Faster R-CNN C4 [31] configuration, using ResNet50 [14] backbone pre-trained on ImageNet [9], together with a linear layer (projection layer) to obtain the object feature representations. We use Detectron2 framework [39] for our implementation. For the part-of-word feature representations, we use the embedding module of the pre-trained BERT [10] "base-uncased" model from the HuggingFace implementation [38]. To get the object proposals for the LSM stage, we train a generic object proposal network, OLN [18]. OLN is trained using only the known classes

Table 1: Comparing mAP and $AP_{50}$ state-of-the-art methods. LocOv outperforms all other methods for Novel objects in the generalized setup while using only 0.6M of image-caption pairs. Training dataset: *ImageNet1k, §COCO captions, †CLIP400M, ‡Conceptual Captions, ⋆Open Images, and cCOCO

| Method | Img-Cap Data Size | Constrained | | | | Generalized | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Novel (17) | | Known (48) | | Novel (17) | | Known (48) | | All (65) | |
| | | AP | $AP_{50}$ | AP | $AP_{50}$ | AP | $AP_{50}$ | AP | $AP_{50}$ | AP | $AP_{50}$ |
| Faster R-CNN | | | - | - | 54.5 | - | - | - | - | - | - |
| SB [2] | - | - | 0.70 | - | 29.7 | - | 0.31 | - | 29.2 | - | 24.9 |
| LAB [2] | | - | 0.27 | - | 21.1 | - | 0.22 | - | 20.8 | - | 18.0 |
| DSES [2] | | - | 0.54 | - | 27.2 | - | 0.27 | - | 26.7 | - | 22.1 |
| DELO [45] | | - | 7.6 | - | 14.0 | - | 3.41 | - | 13.8 | - | 13.0 |
| PL [30] | | - | 10.0 | - | 36.8 | - | 4.12 | - | 35.9 | - | 27.9 |
| STT-ZSD (Ours) | | 0.21 | 0.31 | 33.2 | 53.4 | 0.03 | 0.05 | **33.0** | 53.1 | 24.4 | 39.2 |
| OVR*§c [40] | 0.6M | 14.6 | 27.5 | 26.9 | 46.8 | - | 22.8 | - | 46.0 | 22.8 | 39.9 |
| LocOv *§c (Ours) | | **17.2** | 30.1 | **33.5** | 53.4 | **16.6** | **28.6** | 31.9 | 51.3 | **28.1** | 45.7 |
| XP-Mask‡§⋆c [16] | 5.7M | - | 29.9 | - | 46.8 | - | 27.0 | - | 46.3 | - | 41.2 |
| CLIP (cropped reg)† [13] | 400M | - | - | - | - | - | 26.3 | - | 28.3 | - | 27.8 |
| RegionCLIP†§c [43] | 400.6M | - | **30.8** | - | **55.2** | - | 26.8 | - | 54.8 | - | 47.5 |
| ViLD†c [13] | 400M | - | - | - | - | - | 27.6 | - | **59.5** | - | **51.3** |

on COCO training set. We use all the proposals generated for the training images which have an objectness score higher than 0.7. For our cross-attention model, we use a transformer-based architecture with 6 hidden layers and 8 attention heads trained from scratch. We train our LSM stage with a base learning rate of 0.001, where the learning rate is divided by 10 at 45k and 60k iterations. We use a batch size of 32 and train on 8 GeForce-RTX-2080-Ti GPUs for 90k iterations. For the STT stage, we initialize the weights of the Faster R-CNN and projection layer from the LSM stage, freezing the first two blocks of ResNet50 and the projection layer. For object classes that contain more than one part-of-word representation given BERT embedding module, we consider the average of their vector representation. We use a base learning rate of 0.005 with a 10 times drop at 60k iterations and do early stopping to avoid over-fitting.

### 4.2 Baselines

**OVR**. The main baseline approach is proposed by Zareian *et al.* [40]. We utilize some components proposed in the work including the two-stage design, grounding loss and usage of a cross-attention model. In this work, we propose new components, which simplify and improve the model performance over OVR.
**STT-ZSD**. Our second baseline uses only the Specialized Task Tuning stage. This resembles a zero-shot object detection setting. The model is initialized with ImageNet [9] weights with a trainable projection layer.
**Zero-shot methods**. We compare to some zero-shot object detection approaches which do not include the weak supervision provided by the captions. We compare to three background-aware zero-shot detection methods, introduced by Bansal *et al.* [2], which project features of an object bounding box proposal method to word embeddings. The **SB** method includes a fixed vector for the background

Table 2: Different image regions for the LSM stage. $R^I_{grid}$- grid-regions, $R^I_{box}$- proposed box-regions and $R^I_{ann}$- ground truth box-regions of (k) known or (n) novel objects use during the LSM stage

| Regions | | | Novel (17) | | | Known (48) | | | Generalized | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $R^I_{grid}$ | $R^I_{box}$ | $R^I_{ann}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ |
| 100 | | k+n | 18.2 | 31.6 | 18.2 | 32.5 | 52.7 | 34.0 | 27.9 | 46.0 | 28.8 |
| | | k+n | 16.3 | 28.4 | 15.9 | 32.9 | 53.1 | 34.9 | 27.6 | 45.3 | 28.8 |
| 100 | 100 | | **17.2** | **30.1** | **17.5** | 33.5 | 53.4 | 35.5 | **28.1** | **45.7** | **29.6** |
| 200 | | | 15.5 | 27.1 | 15.4 | 32.2 | 52.1 | 33.9 | 27.1 | 44.5 | 28.2 |
| | 200 | | 13.7 | 25.7 | 12.9 | **34.2** | **53.8** | **36.5** | 27.5 | 43.8 | 29.1 |

class in order to select which bounding boxes to exclude during the object classification, **LAB** uses multiple latent vectors to represent the different variations of the background class, and **DSES** includes more classes than the known set as word embedding to train in a more dense semantic space. **DELO** [45] method uses a generative model and unknown classes to synthesize visual features and uses them while training to increase background confidence. **PL** [30] work deals with the imbalance between positive vs. negative instance ratio by proposing a method that maximizes the margin between foreground and background boxes. **Faster R-CNN**. We also compare with training the classical Faster R-CNN model only using the known classes.

**Open-vocabulary with large data**. We compare our method with recent open-vocabulary models. RegionClip [43] uses the CLIP [29] pre-trained model to produce image-region pseudo labels and train an object detector. CLIP (cropped reg) [13] uses the CLIP pre-trained model on 400M image-caption pairs on object proposals obtained by an object detector trained on known classes. XP-Mask [16] learns a class-agnostic region proposal and segmentation model from the known classes and then uses this model as a teacher to generate pseudo masks for self-training a student model. Finally, we also compare with VILD [13] which uses CLIP soft predictions to distil semantic information and train an object detector.

### 4.3 Results

**COCO dataset.** Table 1 shows the comparison of our method with several zero-shot and open-vocabulary object detection approaches. LocOv outperforms previous zero-shot detection methods, which show weak performance on detecting novel objects. In comparison to OVR, we improve by 2.53 AP, 3.4 $AP_{50}$ for the novel classes and 3.91 AP, 3.92 $AP_{50}$ for the known categories. We observe open-vocabulary methods including OVR and our methods have a trade-off between known and novel class performance. Our method finds a better trade-off as compared to the previous work. It reduces the performance gap on known classes as compared to the Faster R-CNN and improves over the novel classes as compared to all previous works. Our method is competitive with recent state-

(a) Ground Truth (b) STT-ZSD (c) OVR [40] (d) LocOv

Fig. 3: Qualitative results for open-vocabulary object detection on MSCOCO dataset. Novel classes are shown in magenta while known are in green. Methods compared are described in Section 4.2. (Best viewed in color)

of-the-art methods which use more than ∼700 times more image-captions pairs to train, which makes our method data efficient.

Figure 3 shows some qualitative results of our method compared with the STT-ZSD baseline and OVR. Known categories are drawn in green while novel are highlighted in magenta. The columns correspond to the ground truth, STT-ZSD, OVR and our method from left to right. LocOv is able to find novel objects with a high confidence, such as the dogs in the first example, the couch in the second and the umbrella in the third one. We observe that our method sometimes misclassifies objects with plausible ones, such as the case of the chair in the second example which shares a similar appearance to a couch. These examples show a clear improvement of our approach, over the other methods. In the supplementary material we include some examples of our method showing the limitations and main cause of errors of LocOv .

### 4.4 Ablation Experiments

**Localized objects matter.** Table 2 presents the impact of using box- vs grid-region features in the LSM stage. We compare our method using grid-region features $R^I_{grid}$, proposed box-region features $R^I_{box}$, and using box-region features from the known ($k$) or novel ($n$) class annotations $R^I_{ann}$. We find that the combination of grid- and box-regions proves to be best, showing a complementary

Table 3: Ablation study showing the contribution of our proposed consistency-regularization term ($\mathcal{L}_{Cons}$) and usage of BERT text embeddings on COCO validation set. We compared using frozen pretrained weights (fz) of the language model and embedding, fine-tuning (ft) or training from scratch

| $\mathcal{L}_{Cons}$ | BERT Model | BERT Emb. | Novel (17) | | | Known (48) | | | Generalized | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ |
| ✓ | | fz | **17.2** | **30.1** | **17.5** | **33.5** | 53.4 | **35.5** | 28.1 | 45.7 | **29.6** |
| ✓ | fz | fz | 16.7 | 29.7 | 16.7 | 33.4 | **53.5** | **35.5** | **28.2** | **45.9** | 29.5 |
| ✓ | | ft | 16.9 | 29.5 | 16.9 | 33.4 | 53.0 | 35.4 | 28.1 | 45.7 | 29.4 |
| ✓ | | scratch | 16.0 | 28.3 | 16.2 | 30.4 | 49.6 | 31.8 | 25.8 | 42.9 | 26.6 |
| | | fz | 15.4 | 27.9 | 15.2 | 32.2 | 52.1 | 34.1 | 26.3 | 43.6 | 27.3 |

behavior. We also considered two oracle experiments (row 1 and 2) using ground-truth box-region features from both known and novel class annotations instead of proposed box-region features. The best performance is achieved when combined with additional grid regions (row 1). The additional grid-regions help in capturing the background objects beyond the annotated classes while box-regions focus on foreground objects, which improves the image-caption matching.

**Consistency loss and text embedding selection.** Table 3, shows the contribution of our consistency-regularization term. We get an improvement of 1.76 AP by introducing our consistency loss. We compare the performance of using a pre-trained text embedding module vs learning it from scratch, fine-tuning it or considering the complete contextualized language model during the LSM stage in Table 3. Using the pre-trained text embedding, results in a better model.

We find out that using only the embeddings module is sufficient and better than using the complete contextualized BERT language model for the task of object detection. We argue that this is because objects are mostly represented by single word vectors, using simple disentangled text embeddings is better suited for generating object class features. In the supplementary material we include an ablation study showing that both stages of training are necessary and complementary for the success of LocOv .

Table 4 shows the the improvement in performance for each of our contributions. Our baseline method is our implementation of OVR [40]. Both the consistency-regularization together with the inclusion of the box-regions gives the most increment in performance for both novel and known classes. Using only the BERT Embeddings improves the novel class performance although it affects the known classes. Overall we can see that the three contributions are complementary and improve the method for open-vocabulary detection.

## 5 Conclusion

In this work, we proposed an image-caption matching method for open-vocabulary object detection. We introduced a localized matching technique to learn im-

Table 4: Ablation study showing the contributions LocOv . $\mathcal{L}_{Cons}$ = consistency-regularization, $R^I_{box}$ = inclusion of box-regions together with grid-regions, BERT Emb. only.

| $\mathcal{L}_{Cons}$ | $R^I_{box}$ | BERT Emb. | Novel (17) | | | Known (48) | | | Generalized | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ |
| ✓ | ✓ | ✓ | **17.2** | **30.1** | **17.5** | **33.5** | 53.4 | **35.5** | 28.1 | 45.7 | **29.6** |
| ✓ | ✓ | × | 16.7 | 29.7 | 16.7 | 33.4 | **53.5** | **35.5** | **28.2** | **45.9** | 29.5 |
| ✓ | × | ✓ | 15.5 | 27.1 | 15.4 | 32.2 | 52.1 | 33.9 | 27.1 | 44.5 | 28.2 |
| × | ✓ | ✓ | 15.4 | 27.9 | 15.2 | 32.2 | 52.1 | 34.1 | 26.3 | 43.6 | 27.3 |
| × | × | × | 14.3 | 25.6 | 14.4 | 28.1 | 47.8 | 29.3 | 23.7 | 40.9 | 24.5 |

proved labels of novel classes as compared to only using grid features. We also showed that the language embedding model is preferable over a complete language model, and proposed a regularization approach to improve cross-modal learning. In conjunction, these components yield favorable results compared to previous open-vocabulary methods on COCO and VAW benchmarks, particularly considering the much lower amount of necessary data to learn from.

## Acknowledgement

This work was supported by Deutscher Akademischer Austauschdienst - German Academic Exchange Service (DAAD) Research Grants - Doctoral Programmes in Germany, 2019/20; grant number: 57440921.
The Deep Learning Cluster used in this work is partially funded by the German Research Foundation (DFG) - 417962828.

## References

1. Amrani, E., Ben-Ari, R., Shapira, I., Hakim, T., Bronstein, A.: Self-supervised object detection and retrieval using unlabeled videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
2. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
3. Bertasius, G., Torresani, L.: Cobe: Contextualized object embeddings from narrated instructional video. In: Advances in Neural Information Processing Systems (2020)
4. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
5. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

6. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
7. Chen, Z., Ma, L., Luo, W., Wong, K.Y.K.: Weakly-supervised spatio-temporally grounding natural sentence in video. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)
8. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009)
10. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
12. Ging, S., Zolfaghari, M., Pirsiavash, H., Brox, T.: Coot: Cooperative hierarchical transformer for video-text representation learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
13. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=lL3lnMbR4WU
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
15. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020)
16. Huynh, D., Kuen, J., Lin, Z., Gu, J., Elhamifar, E.: Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. arXiv preprint arXiv:2111.12698 (2021)
17. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. In: Advances in Neural Information Processing Systems (2019)
18. Kim, D., Lin, T.Y., Angelova, A., Kweon, I.S., Kuo, W.: Learning open-world object proposals without learning to classify. IEEE Robotics and Automation Letters **7**(2), 5453–5460 (2022)
19. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
20. Kosugi, S., Yamasaki, T., Aizawa, K.: Object-aware instance labeling for weakly supervised object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
21. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision (2020)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision (2014)

23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV) (2014)
24. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: International Conference on Learning Representations (2021)
25. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (2019)
26. Miech, A., Alayrac, J.B., Laptev, I., Sivic, J., Zisserman, A.: Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
27. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (2014)
28. Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Learning to predict visual attributes in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
30. Rahman, S., Khan, S., Barnes, N.: Improved visual-semantic alignment for zero-shot object detection. Proceedings of the AAAI Conference on Artificial Intelligence (2020)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015)
32. Ren, Z., Yu, Z., Yang, X., Liu, M.Y., Lee, Y.J., Schwing, A.G., Kautz, J.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
33. Sadhu, A., Chen, K., Nevatia, R.: Video object grounding using semantic roles in language description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
34. Sohn, K., Zhang, Z., Li, C., Zhang, H., Lee, C., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
35. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations (2019)
36. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)
37. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
38. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C.,

Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2020)
39. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
40. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
41. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
42. Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
43. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. arXiv preprint arXiv:2112.09106 (2021)
44. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
45. Zhu, P., Wang, H., Saligrama, V.: Don't even look once: Synthesizing features for zero-shot detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)