# Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling

Dat Huynh[1]* Jason Kuen[2] Zhe Lin[2] Jiuxiang Gu[2] Ehsan Elhamifar[1]
[1]Northeastern University [2]Adobe Research
[1]{huynh.dat,e.elhamifar}@northeastern.edu [2]{kuen, zlin, jigu}@adobe.com

## Abstract

*Open-vocabulary instance segmentation aims at segmenting novel classes without mask annotations. It is an important step toward reducing laborious human supervision. Most existing works first pretrain a model on captioned images covering many novel classes and then finetune it on limited base classes with mask annotations. However, the high-level textual information learned from caption pretraining alone cannot effectively encode the details required for pixel-wise segmentation. To address this, we propose a cross-modal pseudo-labeling framework, which generates training pseudo masks by aligning word semantics in captions with visual features of object masks in images. Thus, our framework is capable of labeling novel classes in captions via their word semantics to self-train a student model. To account for noises in pseudo masks, we design a robust student model that selectively distills mask knowledge by estimating the mask noise levels, hence mitigating the adverse impact of noisy pseudo masks. By extensive experiments, we show the effectiveness of our framework, where we significantly improve mAP score by 4.5% on MS-COCO and 5.1% on the large-scale Open Images & Conceptual Captions datasets compared to the state-of-the-art.*[1]

## 1. Introduction

Instance segmentation is a crucial yet challenging task of segmenting all objects in an image with applications in autonomous driving, surveillance systems, and medical imaging. Segmentation works have achieved impressive results thanks to advances in training high capacity models with large amounts of mask annotations [1–4]. To be specific, most methods adopt a two-stage object detection architecture [5] for this task by learning an additional mask head to segment objects within box proposals [6–9]. Recent works focus on high-quality mask segmentation by increasing the prediction resolutions using dynamic networks [10, 11] or boundary refinement [12–14]. Despite their success, these works all require costly mask annotations of every class. As a result, it is difficult to scale such systems to hundreds or thousands of classes due to their high mask annotation costs for training. In this work, we aim to significantly reduce the amount of mask supervision by segmenting novel classes using low-cost captioned images.

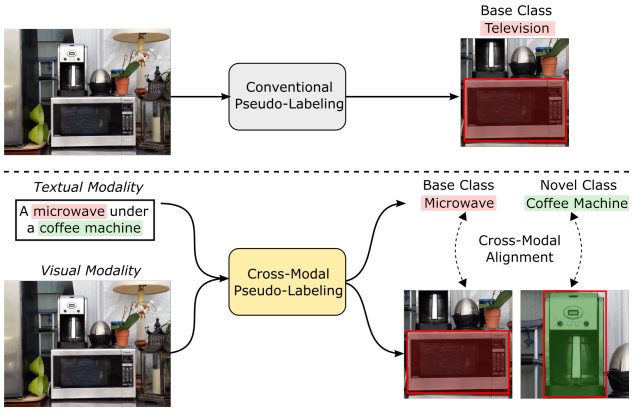


Figure 1. Conventional pseudo-labeling (**top**) only segments objects based on visual modality, which produces incorrect labels and misses novel object classes. Our method (**bottom**) leverages both visual and textual modalities by aligning semantics of caption words with visual features of object masks to correctly label objects and generalize to novel classes without mask annotations.

One of the most popular ways to increase the number of segmentation classes is partially-supervised learning. It utilizes weak image-level [15–17] or box-level [18–23] supervision to segment objects that have no mask annotations, thus lowering the annotation costs. Despite the successes of partially-supervised methods, they can only segment the classes covered by the image/box-level annotation and not a wide general range of novel classes.

Different from previous approaches that are limited to classes with mask annotations, zero-shot instance/semantic segmentation aims to segment novel classes without train-

*This work was done during Dat Huynh's internship at Adobe Research.

[1]Code is available at https://github.com/hbdat/cvpr22_cross_modal_pseudo_labeling.

ing samples via high-level semantic descriptions such as word embeddings. However, current zero-shot approaches on both object detection [24–26] and instance segmentation [27] suffer from low novel-class performances as high-level word embeddings cannot effectively encode fine-grained shape information. To overcome this, the recent OVR [28] work pretrains a visual backbone on captioned images to learn rich visual features. As the backbone of OVR encodes the visual appearances of many novel classes in captions, finetuning it on the detection task significantly improves the performance of novel classes. Despite its effectiveness for detection, we argue that backbone pretraining has limited effects on instance segmentation since mask predictions are ignored and not learned during caption pretraining.

In this paper, we address *instance segmentation of novel classes unknown during training* by directly self-training our model to segment objects in captioned images without any mask annotations. We introduce a robust cross-modal pseudo-labeling framework that aligns textual and visual modalities in captioned images to create caption-driven pseudo masks and generalize to novel classes beyond base classes. Specifically, we train a teacher model on base classes and use this model to select object regions whose visual features are most compatible with the semantics of words in captions. The regions are further segmented into pseudo masks for object words in captions. We then distill pseudo masks into a robust student, which jointly learns segmentation and estimates pseudo-mask noise levels to downweight incorrect teacher predictions. Finally, we evaluate our segmentation performances on MS-COCO and Open Images & Conceptual Captions datasets. We qualitatively demonstrate our generalization ability on truly novel classes, which never appear in most segmentation datasets.

The contributions of this paper are as follows:

• We propose a novel cross-modal pseudo-labeling framework to generate caption-driven pseudo masks and fully utilize captioned images for segmentation training without requiring instance mask annotations.

• Our method is designed to work with novel classes by selecting regions whose visual features are most compatible with the semantics of novel classes and segmenting these regions into pseudo masks to self-train a student model.

• We explicitly capture the reliability of pseudo masks via our robust student model. For pseudo masks with high mask noises, we downweight the loss to avoid error propagation when objects cannot be grounded in images.

• To show the effectiveness of our method, we conduct extensive experiments on MS-COCO and the large-scale Open Images & Conceptual Captions datasets.

## 2. Related Works

**Partially Supervised Learning.** Due to the high cost of mask annotations [29], learning segmentation with weak supervision has attracted strong interest recently. Given bounding box annotations, [15, 20, 21, 30, 31] exploit pixel-wise similarity to infer object masks while [18, 19, 32, 33] learn to share mask knowledge between mask and box supervision to enhance performances. Whereas, [16, 17, 34–38] leverage image-level labels by analyzing classification scores in image regions to estimate object masks. Recently, [39–41] have explored point-wise supervision and learn from only a few background/foreground pixel annotations. Unlabeled images can also be used to improve performances by considering confident predictions as annotations of these images for training [42–48]. However, these works assume certain forms of weak annotations are available for all classes, thus cannot generalize to a wide range of novel classes that may have no annotations at all.

**Zero-Shot Learning**. To generalize toward novel classes without any training annotations, most zero-shot works [49–58] focus on image recognition. Recent works have explored zero-shot object detection by learning to distinguish between background and novel object regions [24, 25], synthesizing unseen class features [26] or using richer textual descriptions [59]. For pixel-level mask prediction, [60–67] perform zero-shot semantic segmentation while [27] tackles the challenging zero-shot instance segmentation task. Since these zero-shot methods only have access to base class annotations, they perform poorly on novel classes. Although [68–70] apply self-training on unlabeled data from novel classes to improve performances, they only address semantic segmentation and cannot distinguish different object instances in an image. Moreover, they make a strong assumption that unlabeled samples always belong to a restricted set of classes known during training.

**Vision-Language Pretraining**, on the other hand, aims to learn from captioned images containing a wide range of classes. Most works focus on learning visual backbones that encode rich visual information from caption-image pairs and finetuning them on downstream tasks. Specifically, [71–75] employ pretrained language models and object detectors to learn visual features well aligned with the embeddings of caption words. Recent works [76, 77] improve training efficiency by removing the need for object detectors and scale to hundreds of millions of samples for substantial performance gains [78]. Moreover, [28] proposes a novel open-vocabulary learning task and shows that pretrained visual features improve not only the detection performance on base classes but also novel classes. However, backbone pretraining alone cannot exploit captioned images for segmentation, as the model is not trained explicitly to segment the objects in captioned images.

**Learning with Noisy Annotations**. Although learning with noisy training samples collected from the web or annotated by machine can also significantly reduce the annotation cost, [79] shows that deep neural networks can easily fit random label noises. Thus, most works address this by regulating the loss function [80–88], denoising training samples [89–91], or utilize additional unlabeled data for regularization [92, 93]. As these methods are not applicable to the segmentation task, [62, 94] propose to capture uncertainty in mask predictions to regulate pixel-wise segmentation loss, thus reducing the impacts of noisy annotations. However, they can only estimate noise from the mask annotations belonging to base classes thus are ineffective for novel classes without mask annotations.

# 3. Robust Cross-Modal Pseudo-Labeling on Captioned Images

This section describes our robust cross-modal pseudo-labeling framework, which utilizes caption-image pairs to produce pseudo masks and self-trains a student model. We first describe the problem setting and then present different components in our framework.

## 3.1. Problem Setting

Let $\mathcal{D}_B = \{(\boldsymbol{I}_m, \mathcal{Y}_m)\}_{m=1}^{N_B}$ be the set of training images and instance annotations for a limited set of base classes $\mathcal{V}_B$. Each image $\boldsymbol{I}_m$ is associated with a set of ground-truth (GT) annotations $\mathcal{Y}_m$, which comprises instance masks and their corresponding object classes. In order to segment novel classes, we leverage additional images $\mathcal{D}_C = \{(\boldsymbol{I}_c, \mathcal{Y}_c)\}_{c=1}^{N_C}$ with only image-level captions. Each image $I_c$ is annotated with a caption from which we can extract a set of object nouns $\mathcal{O}_c \subset \mathcal{Y}_c$ in each caption. Since caption annotations are relatively inexpensive to source, the set of caption classes, $|\mathcal{V}_C|$, is significantly larger than base classes, $|\mathcal{V}_C| \gg |\mathcal{V}_B|$, which is the key ingredient to improve the segmentation of novel classes.

We follow [28] to construct a set of target classes, $\mathcal{V}_T$, without any mask annotations and unknown to the model during training. These classes are merely used as a proxy to evaluate the segmentation performance of novel classes during test time. Our model can recognize a much larger number of novel classes, by using the high-level semantic embeddings $\{\boldsymbol{v}_o\}$, for all object classes $o \in \mathcal{V}_B \cup \mathcal{V}_C \cup \mathcal{V}_T$, from a pretrained `BERT` model [95]. Given the BERT embeddings, we transfer the knowledge from base/caption to target classes via class semantic similarity.

## 3.2. Proposed Method

In this section, we present our proposed cross-modal pseudo-labeling framework for open-vocabulary instance segmentation. For each caption-image pair, we generate pseudo masks by selecting the mask predictions whose visual features are most compatible with semantic embeddings of object words in captions. We first construct a teacher model with an embedding head for classification and a class-agnostic mask head for segmentation. Then, we distill the mask knowledge from teacher predictions and captions into a robust student model which jointly learns from pseudo masks and estimates mask noise levels to downweight unreliable pseudo masks.

### 3.2.1 Designing Teacher Model

To effectively extract mask supervision from captioned images, we first introduce a teacher model, $h$, capable of segmenting novel classes based on the word embeddings of these classes. Following [28], we build upon a two-stage detection framework, `Mask R-CNN` [6]. To be specific, we train a class-agnostic region proposal network, $p$, to select a set of region proposals in each image: $\{\boldsymbol{r}_i\}_{i=1}^{N_R} = p(\boldsymbol{I})$.

Given the region proposals, our goal is to classify them to any classes mentioned in the captions extending beyond base classes. Therefore, we replace the conventional fully-connected layer in the classification head of `Mask R-CNN` with an embedding head $h_{\text{Emb}}$. Here, $h_{\text{Emb}}$ maps the region features into the semantic space of word embeddings. With the embedding head, the score of class $o$ for each region is computed as inner-product between the word embedding of the class and the region's visual feature:

$$\boldsymbol{v}_o^\top h_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{r}}^{\boldsymbol{I}}) \quad \forall \boldsymbol{r} \in p(\boldsymbol{I}), \tag{1}$$

where $\boldsymbol{v}_o$ is the word embedding for class $o$, $\boldsymbol{f}_{\boldsymbol{r}}^{\boldsymbol{I}}$ is the visual feature of region $\boldsymbol{r}$ extracted from the visual backbone using `RoIAlign` [6] and $h_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{r}}^{\boldsymbol{I}})$ is the visual embedding of the region. To simplify the notation, we drop the super-script $\boldsymbol{I}$ in $\boldsymbol{f}_{\boldsymbol{r}}^{\boldsymbol{I}}$ which can be inferred from the context. By learning a joint embedding space between visual features and the word embeddings, the teacher can generalize to novel classes without training samples by measuring the compatibility between visual and textual features. We also define the background embedding to be a fixed zero vector, which has been shown to outperform other variants [28]. Thus, a region proposal is considered background if its class scores are lower than the background score. In addition, we also learn a class-agnostic Mask R-CNN-based head to segment object in each region as, $h_{\text{Mask}}(\boldsymbol{f}_{\boldsymbol{r}})$, where $h_{\text{Mask}}(\cdot)$ is a mask head predicting mask logit scores. To train both embedding and mask heads of the teacher, we adopt the ground-truth loss, $\mathcal{L}_{GT}$, consisting of standard detection and segmentation losses as in [6].

Although the teacher can segment novel classes, it cannot effectively perform this and often miss-classifies novel classes due to their lack of training annotations. To provide additional supervision for novel classes without incurring high annotation costs, we propose a cross-modal

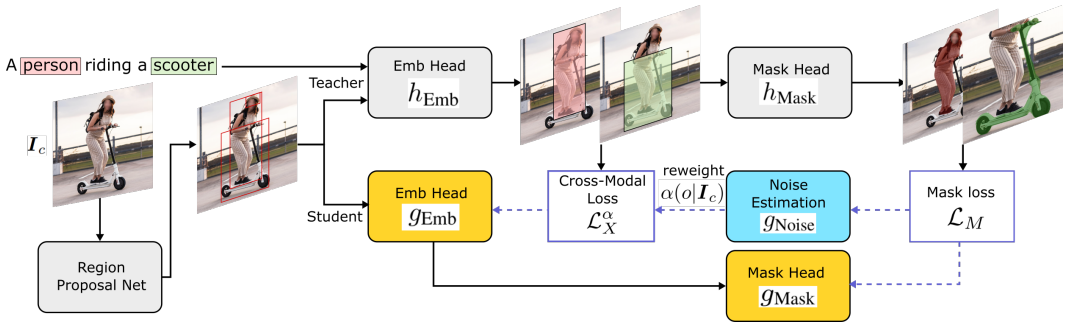


Figure 2. Given an image $\boldsymbol{I}_c$ and the set of objects in captions $\mathcal{O}_c$, we first generate region proposals. We then find the regions that maximize the scores of the teacher embedding head ($h_{\text{Emb}}$) for each object in the caption. We further segment objects within these regions into pseudo masks using the teacher's mask head ($h_{\text{Mask}}$). Finally, the student embedding ($g_{\text{Emb}}$) and mask ($g_{\text{Mask}}$) heads are trained via cross-modal and mask losses, respectively. The cross-modal loss is also reweighted based on the pseudo-mask noise levels learned from our pseudo-mask loss.

pseudo-learning method that uses the semantic information of caption words to guide teacher predictions and generates pseudo masks for self-training a student model.

#### 3.2.2 Cross-Modal Pseudo-Labeling

To boost the teacher's performance in novel classes, we combine the teacher model with caption guidance and explicitly constrain teacher predictions on what objects and where to construct the pseudo masks for training a student model, $g$. We first leverage captions to identify objects in images. For simplicity, we extract object nouns in each caption, $\mathcal{O}_c \subset \mathcal{Y}_c$, as words that are descendants of 'Object' node in the WordNet hierarchy, which is inspired by [3]. To localize these object words in images, we propose a *cross-modal alignment* step that selects the regions whose features are most compatible with the word embeddings of object nouns in captions as following:

$$\boldsymbol{b}_o = \underset{\boldsymbol{r} \in p(\boldsymbol{I}_c)}{\operatorname{argmax}} \big(\boldsymbol{v}_o^\top h_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{r}})\big) \quad \forall o \in \mathcal{O}_c, \tag{2}$$

where the $\boldsymbol{b}_o$ is the *aligned object region* for object $o$ w.r.t. its word embedding $\boldsymbol{v}_o$ and visual embedding from the teacher, $h_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{r}})$. As our pseudo labeling procedure is guided by the word semantics in captions, we specifically search for objects in captions and generalize to novel classes based on their word embeddings. Following recent works on weakly-supervised learning [96, 97], we select the highest confident bounding box for each object to minimize false-positive predictions.

Given the set of aligned object regions, we introduce a cross-modal loss, $\mathcal{L}_X$, which trains the student to identify these regions as their positively-matched caption words:

$$\mathcal{L}_X(\mathcal{Y}_c|\boldsymbol{I}_c; g) = -\sum_{o \in \mathcal{O}_c} \log \frac{\mathrm{e}^{\boldsymbol{v}_o^\top g_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{b}_o})}}{\sum_{w \in \mathcal{V}_C} \mathrm{e}^{\boldsymbol{v}_w^\top g_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{b}_o})}}, \tag{3}$$

where $g_{\text{Emb}}$ is the student embedding head. For each aligned object region $\boldsymbol{b}_o$, the student maximizes its scores of object words in captions and minimizes the scores of other irrelevant words $w$ via Softmax normalization. The information from both word embeddings $\{\boldsymbol{v}_o\}_{o \in \mathcal{O}_c}$ (textual modality) and aligned object regions $\{\boldsymbol{f}_{\boldsymbol{b}_o}\}_{o \in \mathcal{O}_c}$ (visual modality) is distilled into the student embedding head to expand the student's knowledge about the novel classes in captions.

Cross-modal loss works by acting on the student embedding head, but it disregards the mask head that is critical for segmentation. Next, we propose to obtain the pseudo masks from the teacher and estimate the noise levels of such masks. Our method provides supervision for the student mask head, in addition to regulating the cross-modal loss.

#### 3.2.3 Estimating Pseudo-Mask Noises

Given aligned object regions, we turn them into pseudo masks by applying the teacher mask head on these regions:

$$\boldsymbol{M}_o = \mathbb{1}_{\geq 0}[h_{\text{Mask}}(\boldsymbol{f}_{\boldsymbol{b}_o})] \quad \forall o \in \mathcal{O}_c, \tag{4}$$

where $\mathbb{1}_{\geq 0}[\cdot]$ is an indicator function which outputs 1 if a pixel prediction is positive and 0 otherwise to binarize mask predictions. Naively, we can train the student model to mimic the exact pseudo masks at each pixel as:

$$\sum_{o \in \mathcal{O}_c} \sum_{x,y} \mathcal{L}_{\text{BCE}}\big(\boldsymbol{M}_o^{xy} | g_{\text{Mask}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o})\big), \tag{5}$$

where BCE is the binary cross-entropy loss for pixel logit predictions, $\boldsymbol{M}_o^{xy}$ is the pseudo masks at pixel $(x, y)$ and $g_{\text{Mask}}^{xy}$ is the student mask predictions at the pixel. However, not all objects in captions can be correctly detected /segmented due to the errors in teacher predictions, as shown in Figure 3. Thus, minimizing this pixel-wise loss propagates the errors from pseudo masks to the student mask head and degrades its performance. To account for errors in pseudo labels, we propose to estimate the noise level in pseudo masks. Specifically, the student predicts an additional noise value for each pixel in pseudo masks following [62, 94]. We assume that each pixel in a pseudo mask is

corrupted by a Gaussian noise whose variances can be estimated via the visual features of the aligned object region. Thus, we can learn to estimate the pixel-wise noise as:

$$\mathcal{L}_M(\mathcal{Y}_c|\boldsymbol{I}_c, g) = \sum_{o\in\mathcal{O}_c}\sum_{x,y}\mathcal{L}_{\text{BCE}}\big(\boldsymbol{M}_o^{xy}|g_{\text{Mask}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o}) + \epsilon_o^{xy}\big)$$
$$\epsilon_o^{xy} \sim \mathcal{N}\big(0, g_{\text{Noise}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o})\big), \tag{6}$$

where $g_{\text{Noise}}$ is a neural network predicting the noise levels from the visual features of aligned object regions $\boldsymbol{f}_{\boldsymbol{b}_o}$, and $\epsilon_o^{xy}$ is the noise value for the pixel $(x, y)$ of object $o$ sampled from the Gaussian distribution, $\mathcal{N}$, parameterized by $g_{\text{Noise}}$. Pseudo masks with segmentation errors, which are difficult to learn by the student, would drive $g_{\text{noise}}$ to estimate high noise levels to fit these errors. As such, our framework not only trains the student mask head on pseudo masks but also estimates pseudo-mask noise to regulate the training loss and account for possible segmentation errors of the teacher.

With the ability to estimate pseudo-mask noises, we utilize this to improve cross-modal loss in the next section.

#### 3.2.4 Training Robust Student Model

Since both the student and teacher models are unaware of the correct novel object masks due to the lack of annotations, we propose to consider mask noises as a proxy on how reliable the pseudo masks are. We compute the noise level of each pseudo mask as the average of pixel noise: $\sum_{x,y} g_{\text{Noise}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o})/|\boldsymbol{b}_o|$ where $|\boldsymbol{b}_o|$ is the number of pixels in region $\boldsymbol{b}_o$. Then we assign a *reliability score*, $\alpha(o|\boldsymbol{I}_c)$, for each object in captions as the inverse of its average noise level, to indicate the mask reliability:

$$\alpha(o|\boldsymbol{I}_c) = \frac{\eta}{\sum_{x,y} g_{\text{Noise}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o})/|\boldsymbol{b}_o|} \quad \forall o \in \mathcal{O}_c, \tag{7}$$

where $\eta$ is a constant value set to the smallest average noise level across all captioned images[2]. With $\eta$ as the reference, we assign low weights to high-noise predictions while upweighting the clean pseudo masks with low noise levels.

**Objective Function**. Finally, we train a robust student model on datasets of caption and base classes as:

$$\min_{g=\{g_{\text{Emb}}, g_{\text{Mask}}, g_{\text{Noise}}\}} \sum_{c\in\mathcal{D}_C}\Big[\mathcal{L}_M(\mathcal{Y}_c|\boldsymbol{I}_c; g) + \mathcal{L}_X^{\alpha}(\mathcal{Y}_c|\boldsymbol{I}_c; g)\Big]$$
$$+ \sum_{m\in\mathcal{D}_B}\mathcal{L}_{GT}(\mathcal{Y}_m|\boldsymbol{I}_m; g), \tag{8}$$

where $\mathcal{L}_X^{\alpha}$ is the cross-modal loss in Eq. (3) modified to reweight its term as: $\alpha(o|\boldsymbol{I}_c) \times \log \frac{e^{\boldsymbol{v}_o^\top g_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{b}_o})}}{\sum_{w\in\mathcal{V}_C} e^{\boldsymbol{v}_w^\top g_{\text{Emb}}(\boldsymbol{f}_{\boldsymbol{b}_o})}}$ for each object, $o \in \mathcal{O}_c$ . Thus, we effectively downweight the cross-modal loss on noisy predictions to avoid the error propagation from teacher to student.

[2]We determine $\eta$ by training our method on a subset of images and set the smallest average noise level during training to be $\eta$.


Figure 3. Visualization of pseudo mask noise levels and their reliability scores for the objects mentioned in captions.

**Remark 1** *As the student is trained with cross-modal pseudo-labeling that leverages novel-class information from captioned images, it is able to surpass the teacher's performance. This is different from conventional knowledge distillation works, where the student is bounded by the teacher's performance.*

## 4. Experiments

We evaluate our proposed method, which is referred to as XPM for Cross(X)-modal Pseudo Mask, for object detection and instance segmentation on MS-COCO and Open Images & Conceptual Captions datasets. Below, we discuss dataset statistics, evaluation metrics, baselines, and implementation details. We then present and analyze our performances on both base and target classes under various settings. Finally, we demonstrate the importance of each proposed component via ablation study and show how our noise estimation approach compares with other variants.

### 4.1. Experimental Setup

**Datasets**. Following the setup of [28], we perform experiments on MS-COCO [98], which contains 48 base classes with mask annotations and 17 target classes for evaluation. The dataset is partitioned into 107,761 training images with 665,387 mask annotations from base classes and 4,836 testing images consisting of 28,538 and 4,614 mask instances for base and target classes, respectively. For captioned images, we use the entire MS-COCO training set with 118,287 images. Each image is annotated with five captions describing the visually-grounded objects in the image.

To show the effectiveness of our method on large numbers of images and classes, we use large-scale datasets: Open Images [2] with 2.1M instance masks for 300 classes, and Conceptual Captions [99] with 3M captioned images. We propose to split Open Images classes into 200 most common classes as base classes with mask annotations while leaving the remaining 100 rarest classes as target classes unknown to the model during training. Thus, we simulate the real-world setting where the rare classes might be unknown during training.

**Evaluation Metrics.** For both detection and segmentation experiments, we report the mean Average Precision (mAP)

Table 1. Object Detection (mAP) performances trained with bounding-box or mask supervision on base classes in MS-COCO under constrained setting, which outputs either base or target classes, and generalized setting, which must predict all classes. Improvements w.r.t. to other baselines are highlighted in **blue**. * indicates performances reported in [28] while we implement others.

| Method | Bounding Box Supervision | | | | | Instance Mask Supervision | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Constrained | | Generalized | | | Constrained | | Generalized | | |
| | Base | Target | Base | Target | All | Base | Target | Base | Target | All |
| *Zero-Shot Training* | | | | | | | | | | |
| SB* [24] | 29.7 | 0.7 | 29.2 | 0.3 | 24.9 | - | - | - | - | - |
| BA-RPN* [27] | - | 11.4 | 46.5 | 4.8 | 35.6 | - | - | - | - | - |
| *Caption Pretraining with [28]* | | | | | | | | | | |
| OVR [28] | 46.8 | 27.5 | 46.0 | 22.8 | 39.9 | 47.2 | 25.9 | **46.7** | 20.7 | 39.9 |
| SB [24] | 46.9 | 26.9 | 46.3 | 21.2 | 39.7 | 45.9 | 25.7 | 45.3 | 19.6 | 38.6 |
| BA-RPN [27] | 46.8 | 26.0 | 46.2 | 20.7 | 39.5 | 46.0 | 25.0 | 45.5 | 19.3 | 38.7 |
| OVR+OMP [19] | - | - | - | - | - | 34.1 | 16.9 | 33.2 | 10.0 | 27.1 |
| *Pseudo-Labeling* | | | | | | | | | | |
| Soft-Teacher [47] | 47.4 | 18.8 | 47.1 | 12.4 | 38.0 | 46.6 | 16.0 | 46.2 | 10.4 | 36.8 |
| Unbiased-Teacher [48] | **47.5** | 20.5 | **47.2** | 13.8 | 38.4 | 46.6 | 16.8 | 46.1 | 10.8 | 36.9 |
| Cap2Det* [97] | - | - | 20.1 | 20.3 | 20.1 | - | - | - | - | - |
| XPM (Ours) | 46.8 | **29.9**$^{+2.4}$ | 46.3 | **27.0**$^{+4.2}$ | **41.2** | **47.3** | **33.2**$^{+7.3}$ | 46.3 | **29.9**$^{+9.2}$ | **42.0** |

Table 2. Instance Segmentation (mAP) performances in MS-COCO and Open Images & Conceptual Captions datasets.

| Method | MS-COCO | | | | | Open Images & Conceptual Captions | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Constrained | | Generalized | | | Constrained | | Generalized | | |
| | Base | Target | Base | Target | All | Base | Target | Base | Target | All |
| *Caption Pretraining with [28]* | | | | | | | | | | |
| OVR [28] | 42.0 | 20.9 | **41.6** | 17.1 | 35.2 | 52.6 | 23.8 | 45.6 | 17.5 | 36.2 |
| SB [24] | 41.6 | 20.8 | 41.0 | 16.0 | 34.5 | 52.8 | 24.8 | 46.4 | 17.3 | 36.6 |
| BA-RPN [27] | 41.8 | 20.1 | 41.3 | 15.4 | 34.5 | 52.9 | 25.3 | 47.3 | 16.9 | 37.1 |
| OVR+OMP [19] | 31.3 | 14.1 | 30.5 | 8.3 | 24.7 | 52.5 | 24.9 | 47.1 | 16.8 | 36.9 |
| *Pseudo-Labeling* | | | | | | | | | | |
| Soft-Teacher [47] | 41.8 | 14.8 | 41.5 | 9.6 | 33.2 | 52.0 | 25.9 | 46.6 | 17.6 | 36.8 |
| Unbiased-Teacher [48] | 41.8 | 15.1 | 41.4 | 9.8 | 33.1 | 51.7 | 22.2 | 45.3 | 14.5 | 34.9 |
| XPM (Ours) | **42.4** | **24.0**$^{+3.1}$ | 41.5 | **21.6**$^{+4.5}$ | **36.3** | **55.1** | **31.6**$^{+5.7}$ | **49.8** | **22.7**$^{+5.1}$ | **40.7** |

at intersection-over-union (IoU) of 0.5 following conventional zero-shot settings [24, 27, 28]. To analyze the performances on base and target classes, we measure the mAP scores in two settings: i) *constrained setting* where the model is only evaluated on test images belonging to either base classes or target classes; ii) *generalized setting* in which a model is tested jointly on both base and target class images. The latter setting is more challenging as it requires the model to segment target classes and avoid the base-class bias where the model detects target classes as base classes with high confidence.

**Baselines.** We compare with SB [24], which assigns a non-zero background embedding with norm one to predict different background score per bounding box, and open vocabulary object detection OVR [28], which pretrains its embedding space on caption-image pairs. To compare with conventional pseudo-labeling baselines, we adapt Soft-Teacher [47] and Unbiased-Teacher [48], which only use visual modality to construct pseudo labels, by using embedding heads for novel class recognition. In addition, we include the state-of-the-art BA-RPN [27] for zero-shot instance segmentation, which proposes to synchronize background classifier between region proposal network and detection heads to reduce background confusion. We also combine OMP [19] with OVR, which augment the class-agnostic mask head with spatial attention features from embedding head. Finally, to learn from captions images, we compare with Cap2Det [97], which produces pseudo labels for only target and base classes.

**Implementation Details.** To be comparable with [28], we use Mask R-CNN architecture with ResNet50 backbone from maskrcnn-benchmark code base. For training the teacher model, we pretrain the backbone following [28] for 150k iterations on MS-COCO and 200k on Conceptual Captions using 8 V-100 GPUs with the batch size of 32 and the initial learning rate of 0.01. Then we finetune the backbone on segmentation/detection tasks with the batch size of 8 for 90k iterations and the learning rate of 0.001 on both MS-COCO and Open Images datasets to obtain the teacher model. The student is initialized with teacher weights and trained on pseudo and ground-truth labels for an additional 70k iterations. We also downweight the detection loss for background class to 0.2 to improve the recall of target classes, similar to [28]. For the robust student

model, we set $\eta = 0.01$, which is the smallest average noise level estimated offline on 10k captioned images. We use the word embeddings from BERT trained on BookCorpus, and English Wikipedia [95]. For training the noise estimation module, $g_{\text{Noise}}$, we use reparametrization trick [100] to backpropagate gradients through the sampled noise value, $\epsilon$. Moreover, we do not optimize $g_{\text{noise}}$ with respect to $\mathcal{L}_X^\alpha$, which would result in the trivial solution where the student always predicts low-reliability scores to minimize the loss.

### 4.2. Experimental Results

**Object Detection.** We evaluate our method for the object detection task under bounding box or mask supervision of base classes in Table 1 on MS-COCO. Based on the base/target class results in the constrained setting and the generalized setting, we make the following conclusions:

• Although using caption pretraining improves the performance on target classes (under bounding box supervision) over zero-shot training, this strategy does not work as well for mask-level supervision. Since caption-based backbone pretraining [28] can only learn high-level spatially-coarse features of objects but not fine-grained object masks, fine-tuning on mask annotations corrupts the learned backbone and degrades its performances on target classes. This shows the incompatibility between the mask prediction task and the information encoded in the pretrained backbone.

• Soft-Teacher and Unbiased Teacher improve the performance on base classes (under box-level supervision) over using caption pretraining alone. However, as these baselines do not constrain their predictions based on captions, they miss-label novel classes, which propagates teacher error and degrades target-class performances. Although Cap2Det conditions its pseudo labels on captions, these labels come from a limited set of base and target classes. Thus, Cap2Det cannot exploit the useful information from other novel objects in the captions.

• With bounding box supervision, our method (without estimating mask noises) significantly improves target class performances by 2.4% and 4.2% in constrained and generalized settings, respectively. This shows the importance of leveraging captions to improve the pseudo labeling of target classes without annotations. Moreover, with additional mask annotations, we further gain 9.2% in performance on target classes compared to state-of-the-art, which shows the effectiveness of self-training on pseudo masks.

**Instance Segmentation.** To show the effectiveness of XPM, we conduct instances segmentation experiments on both MS-COCO and Open Images datasets. We report the results in Table 2 and conclude that:

• On MS-COCO, different background modeling techniques in SB, BA-RPN have minimal impact on target-class performances when combined with embedding-based caption pretraining. On the other hand, explicitly transferring this knowledge from embedding to mask heads via OMP significantly degrades the performances on base and target classes. This is due to the insufficient amount of base classes and training samples to learn meaningful Object Mask Prior from the small-scale MS-COCO dataset.

• On the large-scale Conceptual Captions and Open Images datasets, both SB and BA-RPN improve target-class segmentation in the constrained setting, as more accurate background models can be learned from the larger number of base classes in Open Images compared to MS-COCO. We observe that conventional pseudo-labeling methods Soft-Teacher, Unbiased Teacher have no significant improvements over caption-pretraining baselines since they cannot utilize textual modality in captioned images to spot novel classes correctly.

• Overall, our method achieves significant performance improvements of at least 4.5% and 5.1% mAP score compared to other baselines in MS-COCO and Open Images datasets, respectively. Moreover, in the Conceptual Captions and Open Images setting, we observe a compound effect – as a result of using a larger number of base classes for training, our teacher model generalizes significantly better on target classes. When labeling Conceptual Captions with the teacher, we benefit from the significant increase in pseudo labels' quantity and quality. Thus, the student obtains strong results on both base and target classes, with a significant gain of 3.6% on all classes.

**Ablation Study.** Figure 6 shows our segmentation improvements compared to the teacher model when introducing different components in our method, on both MS-COCO and Open Images & Conceptual Captions. Adding the cross-modal loss, $\mathcal{L}_X$, significantly improves the segmentation performance over the teacher model, as the student can distill rich knowledge from captioned images. Although the mask loss, $\mathcal{L}_M$, improves target-class performance on MS-COCO, it fails to improve with Conceptual Captions due to noisy web captions. By regulating the cross-modal loss with noise estimation, $\mathcal{L}_X^\alpha$, we gain further improvement on both caption datasets by mitigating the error propagation from teacher to student model.

**Effectiveness of Robust Student.** In Table 3, we experiment with other methods on pseudo-mask noise estimation and loss reweighting. We evaluate Stochastic BCE [62] which learns pixel-wise noise to regulate mask loss, $\mathcal{L}_M$. This method is unable to improve performances as it cannot use mask noise to regulate the cross-modal loss for classification. For the methods that regulate cross-modal loss $\mathcal{L}_X^\alpha$, we consider Class Score which uses class prediction confidences, Pixel Score [101] which estimates mask quality by aggregating pixel-wise prediction confidence, and DropOut Entropy [102] which computes prediction entropy via multiple dropout passes. These

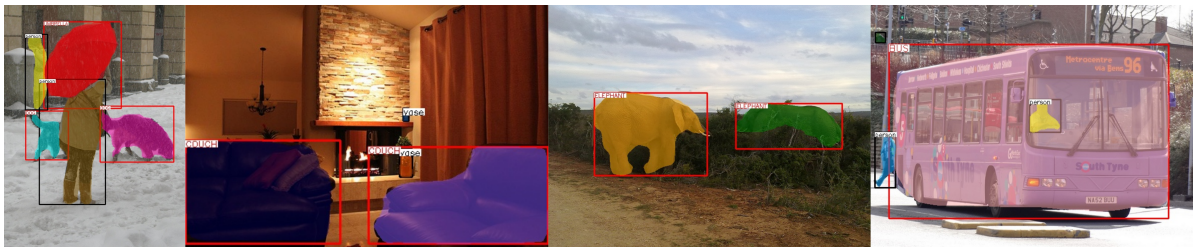

Figure 4. Visualization of our mask predictions for base classes (in **back box**) and target classes (in **red box**) in the generalized setting.

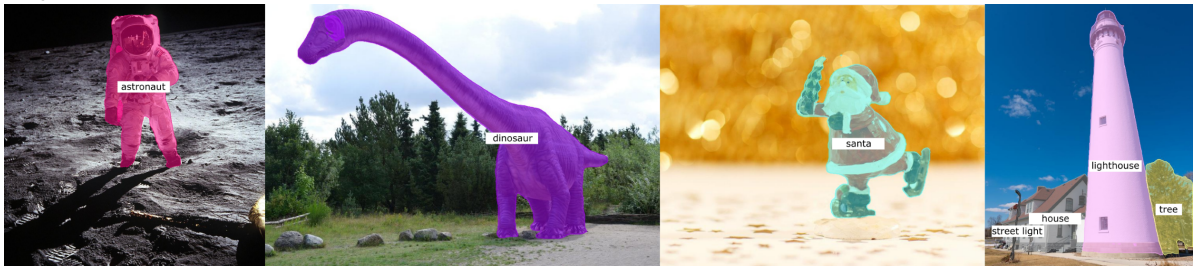


Figure 5. Visualization of our mask predictions for novel classes in the wild with large-scale cross-modal pseudo-labeling.

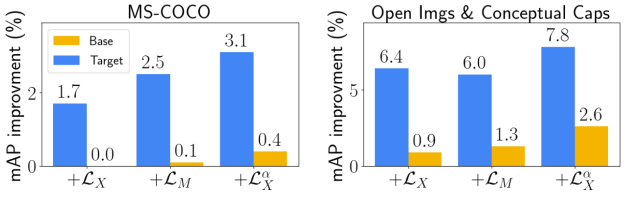


Figure 6. Segmentation improvements w.r.t. the teacher model from adding different proposed components to the student.

Table 3. Segmentation performances of different strategies for noise estimation and loss weighting on Open Images.

| Method | Used on | Base | Target | All |
|---|---|---|---|---|
| `No Noise Estimation` | - | 53.3 | 30.2 | 39.1 |
| `Stochastic BCE` [62] | $\mathcal{L}_M$ | 53.8 | 29.8 | 39.2 |
| `Class Score` | $\mathcal{L}_X^\alpha$ | 54.0 | 28.4 | 38.8 |
| `Pixel Score` [101] | | 53.2 | 30.1 | 38.5 |
| `DropOut Entropy` [102] | | 53.6 | 29.7 | 38.5 |
| `Robust Student (Ours)` | $\mathcal{L}_X^\alpha + \mathcal{L}_M$ | **55.1** | **31.6** | **40.7** |

methods provide no significant improvements, as they are trained on clean annotations of base classes and not adapted to noisy pseudo masks. By learning to estimate the noise levels of pseudo masks and regulating both $\mathcal{L}_X, \mathcal{L}_M^\alpha$, we achieve superior performances compared to `No Noise Estimation`.

**Qualitative Results.** Figure 4 shows the mask predictions of our methods for both base and target classes on MS-COCO. Our method can correctly detect and segment multiple instances of target classes without any ground-truth mask annotations during training. Moreover, our framework maintains strong performances on base classes such that it can correctly segment the base class "bus driver" (the last example) within the target class "bus".

We also visualize the pixel-wise noises for each object in captions in Figure 3. We observe that a good pseudo mask (*e.g.,* 'bear') only has a few noisy pixels along its object boundaries. Whereas, an incorrect pseudo mask (*e.g.,* 'skateboard') contains a large number of noisy pixels that spread over large areas within the bounding boxes.

**Large-Scale Cross-Modal Pseudo-Labeling.** To demonstrate the scalability of our method, we apply cross-modal pseudo-labeling with multiple segmentation datasets (Open Images [2], LVIS [3]), object detection dataset (Objects365 [103]), and caption dataset (Conceptual Captions [99]), to create a high-performance student model. As shown in Figure 5, this strong student, trained with our method, successfully generalized to novel classes such as "astronaut" and "dinosaur", which never appear in most segmentation datasets. Moreover, we can segment the fine details of such truly novel classes without any mask annotations.

## 5. Conclusions

We tackle the problem of open-vocabulary instance segmentation by proposing a robust cross-modal pseudo-labeling framework to provide mask supervision of novel classes in captioned images for training segmentation models. We show the effectiveness of our method on both MS-COCO and Open Images & Conceptual Captions datasets. However, our method might not be suitable for learning with limited base classes as we assume the base classes are sufficiently diverse to enable novel-class generalization.

## Acknowledgement

We would like to thank Ping Hu for his valuable suggestions on implementing the robust student model. This work is partially supported by DARPA (HR00112220001), NSF (IIS-2115110) and ARO (W911NF2110276). Content does not necessarily reflect the position/policy of the Government. No official endorsement should be inferred.

## References


[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[2] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari., "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, 2016. 1, 5, 8

[3] A. Gupta, P. Dollár, and R. B. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 4, 8

[4] S. W. Zamir, A. Arora, A. Gupta, S. H. Khan, G. Sun, F. S. Khan, F. Zhu, L. Shao, G. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," *CVPR Workshops*, 2019. 1

[5] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015. 1

[6] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 3

[7] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[8] Z. M. Chen, X. S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. abs/1904.03582, 2019. 1

[9] H. Huang, C. Wang, P. S. Yu, and C. D. Wang, "Generative dual adversarial network for generalized zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[10] A. Arnab and P. H. S. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[11] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," *European Conference on Computer Vision*, 2020. 1

[12] A. Kirillov, Y. Wu, K. He, and R. B. Girshick, "Pointrend: Image segmentation as rendering," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[13] G. Zhang, X. Lu, J. Tan, J. Li, Z. Zhang, Q. Li, and X. Hu, "Refinemask: Towards high-quality instance segmentation with fine-grained features," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[14] C. Tang, H. Chen, X. Li, J. Li, Z. Zhang, and X. Hu, "Look closer to segment better: Boundary patch refinement for instance segmentation," 2021. 1

[15] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," *Neural Information Processing Systems*, 2019. 1, 2

[16] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[17] A. Arun, C. V. Jawahar, and M. P. Kumar, "Weakly supervised instance segmentation by learning annotation consistent instances," *European Conference on Computer Vision*, 2020. 1, 2

[18] R. Hu, P. Dollár, K. He, T. Darrell, and R. B. Girshick, "Learning to segment every thing," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[19] D. Biertimpel, S. Shkodrani, A. S. Baslamisli, and N. Baka, "Prior to segment: Foreground cues for novel objects in partially supervised instance segmentation," *IEEE International Conference on Computer Vision*, 2021. 1, 2, 6

[20] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[21] Z. Tian, C. Shen, X. Wang, and H. Chen, "Boxinst: High-performance instance segmentation with box annotations," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[22] J. Lee, J. Yi, C. Shin, and S. Yoon, "Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[23] X. Wang, J. Feng, B. Hu, Q. Ding, L. Ran, X. Chen, and W. Liu, "Weakly-supervised instance segmentation via class-agnostic learning with salient images," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[24] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," *European Conference on Computer Vision*, 2018. 2, 6

[25] S. Rahman, S. Khan, and N. Barnes, "Improved visual-semantic alignment for zero-shot object detection," *AAAI Conference on Artificial Intelligence*, 2020. 2

[26] P. Zhu, H. Wang, and V. Saligrama, "Don't even look once: Synthesizing features for zero-shot detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[27] Y. Zheng, J. Wu, Y. Qin, F. Zhang, and L. Cui, "Zero-shot instance segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6

[28] A. Zareian, K. D. Rosa, D. H. Hu, and S. F. Chang, "Open-vocabulary object detection using captions," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 5, 6, 7

[29] A. L. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," *European Conference on Computer Vision*, 2016. 2

[30] A. Khoreva, R. Benenson, J. H. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[31] S. Lan, Z. Yu, C. B. Choy, S. Radhakrishnan, G. Liu, Y. Zhu, L. Davis, and A. Anandkumar, "Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision," *IEEE International Conference on Computer Vision*, 2021. 2

[32] W. Kuo, A. Angelova, J. Malik, and T.-Y. Lin, "Shapemask: Learning to segment novel objects by refining shape priors," *IEEE International Conference on Computer Vision*, 2019. 2

[33] Q. Fan, L. Ke, W. Pei, C.-K. Tang, and Y.-W. Tai, "Commonality-parsing network across shape and appearance for partially supervised instance segmentation," *European Conference on Computer Vision*, 2020. 2

[34] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[35] W. Ge, S. Guo, W. Huang, and M. R. Scott, "Label-penet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation," *IEEE International Conference on Computer Vision*, 2019. 2

[36] P. Zhu, H. Wang, and V. Saligrama, "Generalized zero-shot recognition based on visually semantic embedding," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[37] H. Cholakkal, G. Sun, F. S. Khan, and L. Shao, "Object counting and instance segmentation with image-level supervision," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12 389–12 397, 2019. 2

[38] Y. Shen, L. Cao, Z. Chen, B. Zhang, C. Su, Y. Wu, F. Huang, and R. Ji, "Parallel detection-and-segmentation learning for weakly supervised instance segmentation," *IEEE International Conference on Computer Vision*, 2021. 2

[39] I. H. Laradji, N. Rostamzadeh, P. H. O. Pinheiro, D. Vázquez, and M. W. Schmidt, "Proposal-based instance segmentation with point supervision," *IEEE International Conference on Image Processing*, 2020. 2

[40] B. Cheng, O. Parkhi, and A. Kirillov, "Pointly-supervised instance segmentation," *ArXiv*, 2021. 2

[41] Y. Li, H. Zhao, X. Qi, Y. Chen, L. Qi, L. Wang, Z. Li, J. Sun, and J. Jia, "Fully convolutional networks for panoptic segmentation with point-based supervision," *ArXiv*, 2021. 2

[42] I. Radosavovic, P. Dollár, R. B. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[43] K. Wang, X. Yan, D. Zhang, L. Zhang, and L. Lin, "Towards human-machine cooperation: Self-supervised sample mining for object detection," *European Conference on Computer Vision*, 2018. 2

[44] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *ArXiv*, 2020. 2

[45] J. Li, C. Zhang, P. Zhu, B. Wu, L. Chen, and Q. Hu, "Splmll: Selecting predictable landmarks for multi-label learning," *European Conference on Computer Vision*, 2020. 2

[46] B. Zoph, G. Ghiasi, T. Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. V. Le, "Rethinking pre-training and self-training," *Neural Information Processing Systems*, 2020. 2

[47] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," *IEEE International Conference on Computer Vision*, 2021. 2, 6

[48] Y. C. Liu, C. Y. Ma, Z. He, C. W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," *International Conference on Learning Representations*, 2021. 2, 6

[49] D. Huynh and E. Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[50] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning — the good, the bad and the ugly," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[51] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[52] R. Felix, B. G. V. Kumar, I. D. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," *European Conference on Computer Vision*, 2018. 2

[53] H. Jiang, R. Wang, S. Shan, and X. Chen, "Transferable contrastive network for generalized zero-shot learning," *IEEE International Conference on Computer Vision*, 2019. 2

[54] Y. Atzmon and G. Chechik, "Adaptive confidence smoothing for generalized zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[55] Y. Gong, S. Karanam, Z. Wu, K. Peng, J. Ernst, and P. Doerschuk, "Learning compositional visual concepts with mutual consistency," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[56] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[57] ——, "Compositional zero-shot learning via fine-grained dense feature composition," *Neural Information Processing Systems*, 2020. 2

[58] ——, "Interaction compass: Multi-label zero-shot learning of human-object interactions via spatial relations," *International Conference on Computer Vision*, 2021. 2

[59] Z. Li, L. Yao, X. Zhang, X. Wang, S. S. Kanhere, and H. Zhang, "Zero-shot object detection with textual descriptions," *AAAI Conference on Artificial Intelligence*, 2019. 2

[60] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-vaegan-d2: A feature generating framework for any-shot learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[61] N. Kato, T. Yamasaki, and K. Aizawa, "Zero-shot semantic segmentation via variational mapping," *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 2

[62] P. Hu, S. Sclaroff, and K. Saenko, "Uncertainty-aware learning for zero-shot semantic segmentation," *Neural Information Processing Systems*, 2020. 2, 3, 4, 7, 8

[63] G. Tian, S. Wang, J. Feng, L. Zhou, and Y. Mu, "Cap2seg: Inferring semantic and spatial context from captions for zero-shot image segmentation," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 2

[64] P. Li, Y. Wei, and Y. Yang, "Consistent structural relation learning for zero-shot segmentation," *Neural Information Processing Systems*, 2020. 2

[65] H. Zhang and H. Ding, "Prototypical matching and open set rejection for zero-shot semantic segmentation," *IEEE International Conference on Computer Vision*, 2021. 2

[66] D. Baek, Y. Oh, and B. Ham, "Exploiting a joint embedding space for generalized zero-shot semantic segmentation," *IEEE International Conference on Computer Vision*, 2021. 2

[67] J. Cheng, S. Nandi, P. Natarajan, and W. Abd-Almageed, "Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation," *IEEE International Conference on Computer Vision*, 2021. 2

[68] M. Bucher, T. H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," *Neural Information Processing Systems*, 2019. 2

[69] S. Rahman, S. Khan, and N. Barnes, "Transductive learning for zero-shot object detection," *IEEE International Conference on Computer Vision*, 2019. 2

[70] G. Pastore, F. Cermelli, Y. Xian, M. Mancini, Z. Akata, and B. Caputo, "A closer look at self-training for zero-label semantic segmentation," *Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[71] H. H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *Empirical Methods in Natural Language Processing*, 2019. 2

[72] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Neural Information Processing Systems*, 2019. 2

[73] Y. Li, J. He, X. Zhou, Y. Zhang, and J. Baldridge, "Mapping natural language instructions to mobile ui action sequences," *Annual Meeting of the Association for Computational Linguistics*, 2020. 2

[74] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," *AAAI Conference on Artificial Intelligence*, 2020. 2

[75] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," *European Conference on Computer Vision*, 2020. 2

[76] X. Yuan, Z. L. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[77] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[78] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *International Conference on Machine learning*, 2021. 2

[79] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *International Conference on Learning Representations*, 2017. 3

[80] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari, "Learning with noisy labels," *Neural Information processing Systems*, 2013. 3

[81] E. A. Sanchez, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," *International Conference on Machine learning*, 2019. 3

[82] T. Wang, R. Anwer, M. H. Khan, F. Khan, Y. Pang, L. Shao, and J. Laaksonen, "Deep contextual attention for human-object interaction detection," *IEEE International Conference on Computer Vision*, 2019. 3

[83] X. Zhou, X. Liu, C. Wang, D. Zhai, J. Jiang, and X. Ji, "Learning with noisy labels via sparse regularization," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[84] H. Zhang, X. Xing, and L. Liu, "Dualgraph: A graph-based method for reasoning about label noise," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[85] Y. Xu, L. Zhu, L. Jiang, and Y. Yang, "Faster meta update strategy for noise-robust deep learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[86] D. Ortego, E. Arazo, P. Albert, N. E. O'Connor, and K. McGuinness, "Multi-objective interpolation training for robustness to label noise," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[87] M. Collier, B. Mustafa, E. Kokiopoulou, R. Jenatton, and J. Berent, "Correlated input-dependent label noise in large-scale image classification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[88] Z. Zhu, T. Liu, and Y. Liu, "A second-order approach to learning with instance-dependent label noise," *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[89] A. Veit, N. G. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. J. Belongie, "Learning from noisy large-scale datasets with minimal supervision," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[90] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[91] J. Li, C. Xiong, and S. C. Hoi, "Learning from noisy data with robust representation learning," *IEEE International Conference on Computer Vision*, 2021. 3

[92] Y. Ding, L. Wang, D. Fan, and B. Gong, "A semi-supervised two-stage approach to learning from noisy labels," *IEEE Winter Conference on Applications of Computer Vision*, 2018. 3

[93] J. Li, R. Socher, and S. C. H. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *International Conference on Learning Representations*, 2020. 3

[94] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" 2017. 3, 4

[95] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. 3, 7

[96] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4

[97] K. Ye, M. Zhang, A. Kovashka, W. Li, D. Qin, and J. Berent, "Cap2det: Learning to amplify weak caption supervision for object detection," *IEEE International Conference on Computer Vision*, 2019. 4, 6

[98] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *European Conference on Computer Vision*, 2014. 5

[99] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," *Association for Computational Linguistics*, 2018. 5, 8

[100] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations*. 7

[101] L. Yang, Q. Song, Z. Wang, Z. Liu, S. Xu, and Z. Li, "Quality-aware network for human parsing," *ArXiv*, 2021. 7, 8

[102] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *International Conference on Machine learning*, 2016. 7, 8

[103] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," *IEEE International Conference on Computer Vision*, 2019. 8