Scaling Open-Vocabulary Image Segmentation with Image-Level Labels

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin*

Google Research
{golnazg, xiuyegu, yincui}@google.com tsungyil@nvidia.com

Abstract We design an open-vocabulary image segmentation model to organize an image into meaningful regions indicated by arbitrary texts. Recent works (CLIP and ALIGN), despite attaining impressive open-vocabulary classification accuracy with image-level caption labels, are unable to segment visual concepts with pixels. We argue that these models miss an important step of visual grouping, which organizes pixels into groups before learning visual-semantic alignments. We propose OpenSeg to address the above issue while still making use of scalable image-level supervision of captions. First, it learns to propose segmentation masks for possible organizations. Then it learns visual-semantic alignments by aligning each word in a caption to one or a few predicted masks. We find the mask representations are the key to support learning image segmentation from captions, making it possible to scale up the dataset and vocabulary sizes. OpenSeg significantly outperforms the recent open-vocabulary method of LSeg by +19.9 mIoU on PASCAL dataset, thanks to its scalability.

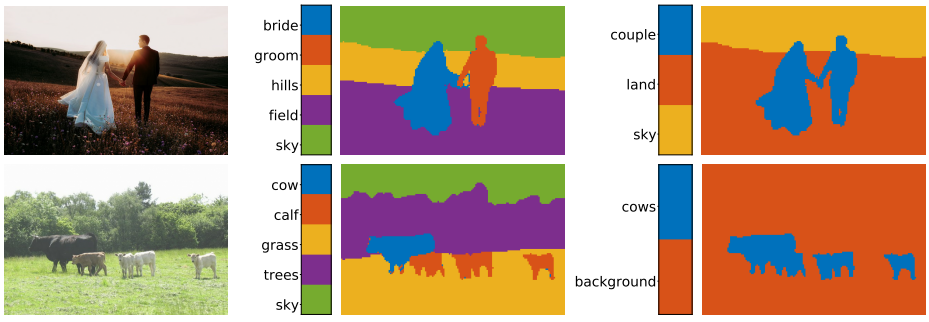


Figure 1. Examples of image segmentation with arbitrary text queries. We propose a model, called **OpenSeg**, that can organize pixels into meaningful regions indicated by texts. In contrast to segmentation models trained with close-vocabulary categories, OpenSeg can handle arbitrary text queries. For example, the model segments out a region for 'couple' and two regions for 'bride' and 'groom'.

* Work done while at Google.

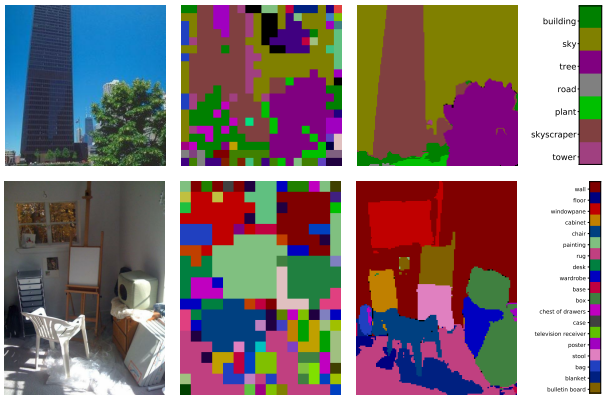


Figure 2. ALIGN (middle) can only roughly localize text queries onto the image. In contrast, OpenSeg (right) can localize visual concepts with accurate segmentation. Moreover, ALIGN predicts more false positives not present in the image.

1 Introduction

Image segmentation is an important step to organize an image into a small number of regions in order to understand *"what"* and *"where"* are in an image. Each region represents a semantically meaningful entity, which can be a thing (*e.g.*, a chair) or stuff (*e.g.*, floor). Language is a natural interface to describe what is in an image. However, semantic segmentation algorithms often only learn with closed-set categories, and thus are unable to recognize concepts outside labeled datasets. Figure 1 shows examples of image segmentation driven by language. The segmentation model takes text queries as inputs and produces segmented regions accordingly. In this work, we aim to learn open-vocabulary models which can segment an image and indicate regions with arbitrary text queries.

Recently, CLIP [40] and ALIGN [23] learn with billion-scale image-text training examples to understand *"what"* are in an image with arbitrary text queries. These models demonstrate impressive results when directly evaluated on downstream image-text retrieval or classification tasks. However, localizing text queries to understand *"where"* these visual concepts are in an image is still challenging. For example, Figure 2 shows the segmentation predictions of a pre-trained ALIGN [23] model using class activation maps [60].

We argue that what is missing in these state-of-the-art open-vocabulary classification models are mid-level representations from visual groupings [48], which organize an image into a small set of segmentation masks. Furthermore, visual-semantic alignments should perform after grouping to align texts to segmentation regions. However, these models represent an image with a single feature vector, inevitably losing much location information.

Recently, Li *et al.* [29] introduce an open-vocabulary segmentation method using pre-trained CLIP [40] text-encoders. It trains an image encoder to predict pixel embedding aligned with the text embedding of its pixel label.

However, the issue with this approach is in the scalability of training data. It is costly to annotate pixel-wise class labels, and thus requires generalization to unseen visual concepts from limited class labels. We show that the visual-semantic alignments of image segmentation can be learned from scalable image caption labels.

In this work, we represent an image with a set of segmentation masks and their features. We implement a class-agnostic segmentation module with region-to-image cross-attention [8,46,10] and train it with class-agnostic segmentation masks. In contrast to the works using similar architectures [46,10], we do not predict the "no object" label $\varnothing$ to indicate if a predicted mask is a valid group of pixels. Considering the training data is only annotated with one possible organization of an image, we allow our model to predict other possible organizations beyond the annotations present in the training data.

Next, we learn visual-semantic alignments based on the predicted masks, which provide two major benefits in training. First, we perform mask-based feature pooling to aggregate pixels inside the predicted mask to generate location-aware region features. Second, the small number of predicted masks makes it easier to learn weakly-supervised alignments between regions to words in an image caption. The ability to learn from weak labels is important for scaling up training data and increasing vocabulary sizes. We call our method **OpenSeg**, standing for open-vocabulary image segmentation.

To evaluate our method, we measure performances on holdout image segmentation datasets. We want to promote the framework where the model is trained with a large scale supervised/weakly-supervised data to learn *generalist* models transferable to other datasets. Such a framework has been recently introduced for image classification [40,23] and object detection [58,18]. To our knowledge, OpenSeg is the first work in image segmentation to demonstrate zero-shot transfer results across datasets using language. This is in stark contrast to the existing evaluation protocol which measures performances of *specialist* models trained and tested using limited labeled data from the same dataset distribution.

In our experiments, we train the mask prediction model using class-agnostic mask annotations in the Panoptic COCO dataset [26]. We show that the model can generalize well to other datasets, reaching superior performances compared with prior works on segmentation proposals [3,33]. Then, we report mean intersection over union (mIoU) metrics for measuring both localization quality and accuracy of open-vocabulary semantic recognition. We compare OpenSeg to the recent open-vocabulary method of LSeg. Thanks to the scalability of OpenSeg, our best model significantly outperforms strongest LSeg model by 19.9 on PASCAL dataset. We also compare OpenSeg to a version of LSeg implemented in our framework, trained on a larger semantic segmentation dataset of COCO (LSeg+). OpenSeg with ResNet-101 backbone outperforms LSeg+ models with similar backbone by 2.7 mIoU on PASCAL-Context (459 categories) and 1.9 mIoU on ADE-20k (847 categories). OpenSeg achieves this improvement mainly because of its ability to make use of image caption data which enables us to train it on a larger set of vocabularies and also a larger set of training examples.

2 Related Work

Grouping for visual recognition: Grouping has been a core research area in mid-level visual representations. The importance of grouping for human perception was pointed out almost a hundred years ago [48]. In machine perception, early works [11,42] group pixels based on local affinities. Arbelaez *et al*. [2] find contour detection and multiscale information helpful to generate segmentation and use it to predict object candidates [3]. COB [33] improves the efficiency and performance by leveraging deep nets. These mid-level region representations are then used for semantic segmentation [33] and object detection [45]. Recently, Qi *et al*. [39] propose to segment all visual entities without considering semantic labels and show generalization to unseen domains. In contrast to [39], our work not only predicts segmentation, but also understands the semantics of segmented regions by open-vocabulary visual-semantic alignments.

Fully-supervised segmentation: To understand the semantics of pixels, several datasets have been developed with an increasing number of images and categories [13,35,5,7]. Models trained on these datasets can only learn to recognize the pre-defined classes, which are at most in the order of hundreds for standard benchmarks. Also, the classes across datasets are not transferable. MSeg [28] points out the ambiguity of class definitions, and manually resolves it to learn a transferable model across datasets. But the model still can not transfer to new visual concepts not present in the dataset. OpenSeg overcomes these drawbacks.

Semantic segmentation with less supervisions: Weakly-supervised semantic segmentation trains with image-level labels [36,47,24,31,52], of which refining CAMs [60] is a popular techniques. Our model also adopts weak image-caption supervision, and it is different in that it has access to a set of class-agnostic segmentation annotations. Furthermore, it can transfer to arbitrary classes while these methods can not. Zero-shot semantic segmentation methods [6,49,20,30,4] aim to segment images with unseen visual concepts using language embeddings. These approaches learn with pixel-wise class labels which are expensive to scale up due to the long-tailed nature mentioned in the previous paragraph. In contrast, we leverage cheap image caption data that covers a wide range of concepts, to achieve better and more practical performance on arbitrary categories. In addition, we evaluate on datasets with much larger number of categories to verify the zero-shot transfer capability.

Open-vocabulary segmentation: Open-vocabulary segmentation aims to overcome the limitation of closed-set vocabulary in previous segmentation works. Zhao *et al*. [59] is the pioneering work that learns a joint pixel and word concept embedding space; however, its vocabulary and knowledge is limited to WordNet and can not take arbitrary texts as input.

In a recent work, Li *et al*. [29] train an image encoder to encode pixel embeddings and use CLIP [40] text embeddings as the per-pixel classifier. Both Zhao *et al*. [59] and Li *et al*. [29] need per-pixel semantic supervision which is expensive to scale up. On the contrary, OpenSeg makes use of cheap image-level supervision such as captions, which allows scaling up training data. There are multiple

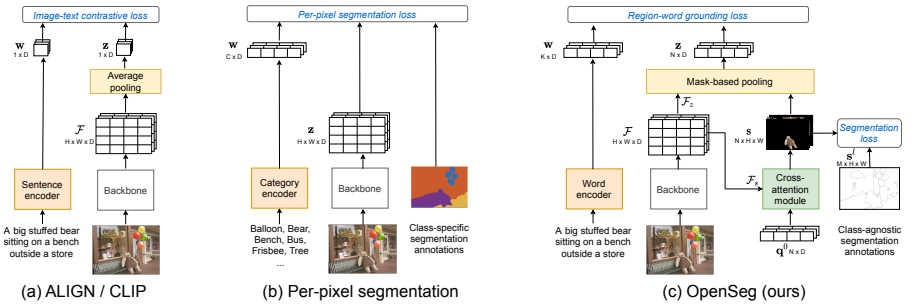


Figure 3. An overview of our approach. We compare OpenSeg with ALIGN / CLIP [23,40] and per-pixel segmentation models such as LSeg [29]. The major differences are in the image and text representations $\mathbf{z}$ and $\mathbf{w}$. ALIGN / CLIP has $\mathbf{z} \in \mathbb{R}^{1\times D}$, losing location information. Per-pixel segmentation represents an image with $\mathbf{z} \in \mathbb{R}^{H\times W\times D}$, requiring class-specific mask annotations for training. OpenSeg represents an image with a set of N segmentation regions $\mathbf{z} \in \mathbb{R}^{N\times D}$, facilitating weakly-supervised learning using captions.

works concurrently developed with OpenSeg: GroupVit [51] learn segmentation masks from text supervision. Zabari and Hoshen [57] use model interpretability to obtain pixel-level pseudo-labels from CLIP to supervise single-image segmentation methods; it's different from all other works as it does not need any training images, but the method is slow. Zhou *et al.* [62] adapt CLIP for segmentation, and use pseudo per-pixel labels and self-training to boost the performance; similar to Li *et al.* [29], it utilizes per-pixel semantic supervision. Xu *et al.* [53] first generate mask proposals, and then leverage CLIP for classification of the proposals. In contrast, we learn visual-semantic alignment from image captions, which is no longer limited by image classification models (*e.g.*, CLIP).

Visual grounding: Image captioning and image text datasets [37,9,27] enable research on the interplay of captions and grounded visual concepts [14,41,15,19,25]. However, these methods often rely on an object detector to predict object bounding boxes for grounding. Therefore, they are not able to handle stuff and can not generate a single segmentation map for everything. Our method also uses captions as semantically-rich supervision. We draw inspiration from these works and expand the model's ability to ground visual concepts of both things and stuff to pixels with our mask representations.

Referring image segmentation: The goal of this task is to compute a binary mask localizing a referring expression. Since there are multiple supervised datasets (*e.g.*, RefCOCO [56]) for this task, previously developed methods are usually fully supervised [21,55,12,22,54]. As a result, the training data for these methods are not scalable.

3 Method

Figure 3 shows an overview of our approach. In contrast to approaches that represent an image with a vector $\mathcal{Z} \in \mathbb{R}^{1\times D}$ or a feature map $\mathcal{Z} \in \mathbb{R}^{H\times W\times D}$, OpenSeg represents an image with N proposal masks with their features $\mathcal{Z} \in \mathbb{R}^{N\times D}$. Our mask representations support learning precise image segmentation with image captions by weakly-supervised learning. In Section 3.1, we describe the learning of predicting mask proposals from an image. In Section 3.2, we describe the feature representations of proposal and the learning of region-word alignments. In the following sections, We use a bold symbol to indicate an array of elements $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, where the first dimension indicates the number of elements.

3.1 Learning Segmentation Masks

We design a model architecture which consists of a feature pyramid network (FPN) [32] for multi-scale feature extraction and a cross-attention module for segmentation region proposal. We fuse FPN features into P_2 resolution as described in [17] to generate image features $\mathcal{F}$. From $\mathcal{F}$, we obtain $\mathcal{F}_s \in \mathbb{R}^{H\times W\times D}$ by convolution and fc layers. Then we augment image features by adding learnable position embeddings PE: $\mathcal{F}_s^{PE} = \mathcal{F}_s + PE$. We use a cross-attention module taking inputs as $\mathcal{F}_s^{PE}$ and a randomly initialized queries $\mathbf{q}^0 \in \mathbb{R}^{N\times D}$ to generate mask queries $\mathbf{q} \in \mathbb{R}^{N\times D}$. Then, we compute the dot product of mask queries and position-augmented image features to predict masks $\mathbf{s} = Sigmoid(dot(\mathbf{q}, \mathcal{F}_s^{PE})) \in \mathbb{R}^{N\times H\times W}$. This architecture is conceptually similar to Max-deeplab [46] and MaskFormer [10]. The details of the architecture are in Appendix C.

We compute Dice coefficient [34] between predicted masks $\mathbf{s}$ and class-agnostic labeled masks $\mathbf{s}^l \in \mathbb{R}^{M\times H\times W}$ and maximize the Dice coefficient of the best matched mask for each labeled mask.

$$\mathcal{L}_{\mathcal{S}} = \frac{1}{M}\sum_{j=1}^{M}(1 - \max_i Dice(s_i, s_j^l)) \tag{1}$$

Typically, $N > M$ for each training image. Therefore, a subset of proposal masks are optimized to best match labeled masks. The rest of proposals can still segment out unlabeled regions without being penalized. One predicted mask may match to multiple labeled masks in the early training stage when their overlaps are low. But this does not prevent learning masks that highly overlap with labeled masks in the latter training stage.

3.2 Visual-Semantic Alignment with Masks

We use a pair of image I_b and caption C_b to learn visual-semantic alignments. We break I_b into regions (Section 3.1) and C_b into words by extracting list of nouns and adjectives from the caption. We randomly drop each word with

the probability of $1 - kp$, where kp is the keep probability of words extracted from captions. We generate image features $\mathcal{F}_z$ using the same architecture as $\mathcal{F}_s$. For each region, we compute its feature by pooling image features with the mask $\mathbf{z}[n, d] = \sum_{ij} \mathbf{s}[n, i, j] \cdot \mathcal{F}_z[i, j, d]$. We feed each word to a pre-trained text encoder to compute the word feature w.

We follow the grounding loss in prior works [19,58] to learn region-word alignments. We first define the notation for Softmax on an array $\mathbf{x}$ to get the normalized score at the i-th element:

$$\sigma(\mathbf{x})_i = \frac{e^{x_i/\tau}}{\sum_j e^{x_j/\tau}} \tag{2}$$

where τ is a learnable scalar for the temperature. The similarity score of a region i and a word j is defined by its cosine similarity $\langle z_i, w_j \rangle = \frac{z_i \cdot w_j}{\|z_i\| \|w_j\|}$. Then we define the similarity of all regions $\mathbf{z}$ to a word w_j as: $g(\mathbf{z}, w_j) = [\langle z_1, w_j \rangle, ..., \langle z_N, w_j \rangle] \in \mathbb{R}^{N \times 1}$. We compute the similarity of an image I_b and its caption C_b by:

$$G(I_b, C_b) = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=1}^{N} \sigma(g(\mathbf{z}, w_j))_i \cdot \langle z_i, w_j \rangle \tag{3}$$

The above similarity function encourages each word to be grounded to one or a few regions. Also, it avoids penalizing regions that can not find any similar word. Next, a grounding loss is defined for a given mini-batch B, where each example contains an image-caption pair. We define the similarity scores of all images in a batch $\mathbf{I}$ to a caption C_b by $G(\mathbf{I}, C_b) = [G(I_1, C_b), ..., G(I_{|B|}, C_b)] \in \mathbb{R}^{|B| \times 1}$ and similarly $G(I_b, \mathbf{C}) = [G(I_b, C_1), ..., G(I_b, C_{|B|})] \in \mathbb{R}^{|B| \times 1}$. The grounding loss aims at maximizing the normalized score of a labeled image-caption pair $\langle I_b, C_b \rangle$ over all images and all captions in a mini-batch.

$$\mathcal{L}_{\mathcal{G}} = -\frac{1}{|B|} \sum_{b=1}^{|B|} \Big(\log \sigma\big(G(\mathbf{I}, C_b)\big)_b + \log \sigma\big(G(I_b, \mathbf{C})\big)_b \Big) \tag{4}$$

To train OpenSeg, we simply sum the two losses with a weight α:

$$\mathcal{L} = \mathcal{L}_{\mathcal{G}} + \alpha \mathcal{L}_{\mathcal{S}} \tag{5}$$

When setting $\alpha = 0$, the model learns without labeled class-agnostic segmentation, and thus needs to induce mask predictions with the visual-semantic grounding loss. We find this setting leads to a poor performance, suggesting class-agnostic mask annotations are critical for learning mask predictions.

3.3 Learning from Caption Only Data

Since annotating images with segmentation is expensive, to scale up the training data we need to learn from images with only caption annotations. We follow MuST [17] and first train a teacher model on a segmentation dataset with only the segmentation loss $\mathcal{L}_{\mathcal{S}}$. Then we annotate a large image-text dataset with pseudo segmentation labels using the teacher model. Lastly, the OpenSeg model is trained with a mix of human and pseudo labels.

3.4 Inference

Up to this point, we learn a vision model that predicts segmentation masks $\mathbf{s} \in \mathbb{R}^{N \times H \times W}$ and corresponding features $\mathbf{z} \in \mathbb{R}^{N \times D}$. Given an evaluation segmentation dataset, we encode its categories using the text encoder. If a category is defined by more than one word, we simply include all word embeddings for that category. We obtain K word embeddings $\mathbf{w} \in \mathbb{R}^{K \times D}$ representing all categories. The region logits are obtained by taking the cosine similarity between words and regions $\langle \mathbf{w}, \mathbf{z} \rangle \in \mathbb{R}^{K \times N}$. We multiply the region logits and segmentation masks to obtain segmentation logits at each pixel $\mathbf{y} = \langle \mathbf{w}, \mathbf{z} \rangle \cdot \mathbf{s} \in \mathbb{R}^{K \times H \times W}$. Then the category prediction at each pixel is an argmax of segmentation logits along the word dimension:

$$pred[i, j] = \underset{k}{\operatorname{argmax}}\, \mathbf{y}[k, i, j] \tag{6}$$

4 Experiments

4.1 Experimental Settings

Architecture. We use EfficientNet-B7 [44] (and ResNet101 in Table 2) as the backbone architecture and employ FPN [32] for multi-scale feature fusion. We use pyramid levels from P_2 to P_5 with feature dimension 640, upsample all feature levels to P_2, and then merge them by a sum operation to obtain $\mathcal{F}$. To compute $\mathcal{F}_z$ and $\mathcal{F}_s$, we apply a fc layer followed by 3 layers of 3×3 convolutions with 640 channels after $\mathcal{F}$. For text encoder we use the frozen pre-trained BERT-Large model in ALIGN [23].

Training Parameters. All models are trained with an image size of 640×640. We apply multi-scale jittering with a random scale between $[0.8, 1.2]$ (*i.e.*, small scale jittering in [16]). The weight decay is set to 1e-05 and we use a learning rate 0.005 with the cosine learning rate schedule. Unless otherwise mentioned, we initialize the backbone of the model from the ALIGN checkpoint [23]. We train OpenSeg on COCO dataset for 30k steps. For training on COCO and Localized Narrative datasets, we sample examples from the datasets with equal probability and we train the model for 60k steps. We set *kp* (keep probability of words extracted from captions) to 0.5. We train models with global batch size of 1024 and local batch size of 16 (we have 64 Cloud TPU v3 cores). Unless otherwise stated, for each core we compute the loss over the local batch of examples (See Appendix F for the comparison between sync and unsync contrastive loss over the cores and also comparison of training with smaller batch sizes).

Training Datasets

COCO: We use the panoptic segmentation [26] and caption [9] annotations in the 2017 splits which include 118k/5k train/val images. We utilize the panoptic segmentation annotations in a class-agnostic manner. When evaluating on COCO Panoptic, we treat it as a semantic segmentation dataset and our model only predicts the semantic class for each pixel.

Table 1. Recall of segmentation mask proposals on COCO and PASCAL-Context datasets. All methods use 128 proposals.

	COCO			PASCAL Context-59		
	R50	R70	R90	R50	R70	R90
MCG [3]	41.1	21.4	4.6	57.8	31.7	8.7
COB [33]	46.0	24.8	4.9	62.9	37.6	12.1
OpenSeg	**68.9**	**48.1**	**16.9**	**84.5**	**65.1**	**29.1**

Localized Narrative (Loc. Narr.): Localized Narrative [38] contains detailed natural language descriptions along with mouse traces for multiple datasets (COCO, Flickr, Open Images, ADE20k). We don't train on the ADE20k portion to keep its image distribution unseen. The remaining 652k images are used for training.

Evaluation Datasets

PASCAL Context: PASCAL Context [35] includes per-pixel segmentation annotations of object and stuff on 5k/5k train/val images from various indoor and outdoor senses. The full version (`PC-459`) includes 459 classes. The version with the most frequent 59 classes (`PC-59`) is widely used in the existing literature.

PASCAL VOC: PASCAL VOC 2012 [13] includes 20 object classes and a background class with 1.5k/1.5k train/val images. Since the text "background" is ambiguous, we assign the background class to the pixels predicted as `PC-59` categories that are not in PASCAL VOC.

ADE20k: ADE20k [61] includes 20k/2k train/val images with segmentation annotations and covers a wide variety of indoor and outdoor scenes. The full version has annotations in an open-vocabulary setting and includes 2693 object and stuff classes. We follow [10] and evaluate on the version with 847 classes (`A-847`). We also test on the widely-used version with 150 frequent categories (`A-150`).

4.2 Predicting Masks Across Datasets

We train the segmentation proposal model on COCO and evaluate on COCO and `PC-59` with recalls at IoU 50%, 70%, and 90% as metrics. Table 1 shows performance comparisons with MCG [3] and COB [33] using their pre-computed proposals. OpenSeg shows significantly superior performances. We perform additional cross-dataset evaluation using datasets in MSeg [28] in Appendix D. Figure 4 shows 6 manually selected proposals to demonstrate our model can organize images into semantically meaningful regions. Particularly, the underwater scene is not present in our training dataset COCO, but the model can still organize pixels into regions for ocean, coral, diver, goggles, *etc*. The full 128 proposals are included in Appendix E.

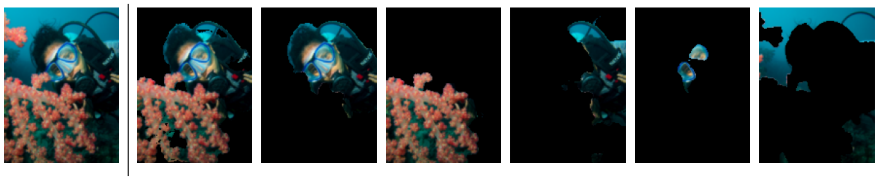


Figure 4. Examples of predicted segmentation masks in an unseen scene. OpenSeg is able to segment an image into meaningful regions. These regions may be overlapping and indicate concepts of foreground (diver and coral) *vs.* background (ocean), and whole (diver) *vs.* parts (scuba and goggles). Notably, OpenSeg is trained on COCO which does not include underwater scenes.

4.3 Open-vocabulary Image Segmentation

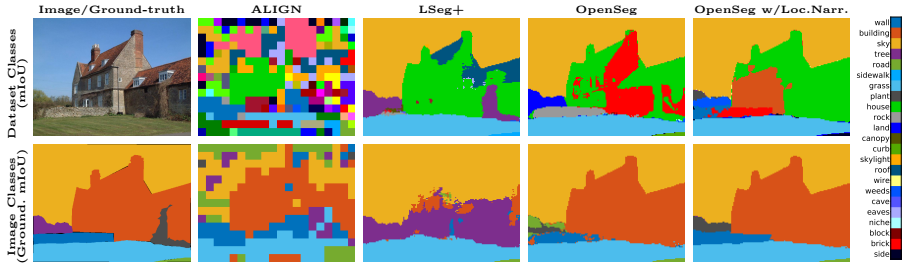
In this section, we first describe open-vocabulary baselines and our evaluation metrics. Then we discuss the experimental results with our open-vocabulary baselines and state-of-the-art open-vocabulary and zero-shot methods.

ALIGN baseline: Although ALIGN [23] is trained for open-vocabulary classification, it can still roughly localize objects and stuff with arbitrary text queries (see Figure 2). Since we initialize the backbone of OpenSeg from ALIGN's pre-trained checkpoint, we use ALIGN as a baseline. We follow the CAM [60] method for segmentation prediction. We compute the activation map before the average pooling layer of the image encoder. Then for each spatial location we compute its cosine similarity with the text embeddings of all input categories. We assign the class with the highest similarity to each location.

LSeg baseline: Recently, [29] introduce an open-vocabulary segmentation method which trains an image encoder to encode pixel embeddings and use CLIP [40] text embeddings as the per-pixel classifier. Figure 3(b) illustrates the model of this approach. For a fair comparison, we also construct LSeg in our codebase as follows. We add FPN and introduce a high resolution map in the same approach in Section 4.1. We embed class names into text embeddings using ALIGN [23] text-encoder and use them as per-pixel classifiers. We fine-tune the pre-trained image encoder and FPN layers on COCO dataset using a per-pixel cross-entropy loss to align pixel embeddings with text embeddings. We call this model LSeg+.

ALIGN w/proposal baseline: The ALIGN, LSeg and LSeg+ baselines are methods that perform visual-semantic alignments without explicit visual grouping. Since our method uses visual grouping, we also compare our method to ALIGN w/proposal baseline which leverage proposals generated by OpenSeg at inference. We use the ALIGN model to classify each proposal and then similarly to OpenSeg we aggregate all proposals to compute the final segmentation map.

Evaluation metrics: We use two metrics, *mIoU* and *Grounding mIoU*, for evaluation. Both metrics are calculated using the standard mIoU formula [13] and only differ in the text queries for each image. The mIoU is commonly used in literature. It measures the performance of image segmentation with fixed text



	COCO Train			mIoU					Grounding mIoU				
	label	mask	cap.	A-847	PC-459	A-150	PC-59	COCO	A-847	PC-459	A-150	PC-59	COCO
ALIGN	✗	✗	✗	4.8	3.6	9.7	18.5	15.6	17.8	21.8	25.7	34.2	28.2
ALIGN w/proposal	✗	✓	✗	5.8	4.8	12.9	22.4	17.9	17.3	19.7	25.3	32.0	23.6
LSeg+	✓	✓	✗	3.8	7.8	18.0	**46.5**	55.1	10.5	17.1	30.8	56.7	60.8
OpenSeg	✗	✓	✓	6.3	9.0	21.1	42.1	36.1	21.8	32.1	41.0	57.2	48.2
OpenSeg w/L. Narr.	✗	✓	✓	**6.8**	**11.2**	**24.8**	45.9	38.1	**25.4**	**39.0**	**45.5**	**61.5**	48.2

Figure 5. (Bottom) The mIoU and Grounding mIoU results of ALIGN, ALIGN w/proposal, LSeg+, and OpenSeg. (Top) Segmentation predictions on an image from the ADE20k (847 categories). (First row) Predictions with all 847 classes as text queries. (Second row) Predictions with only classes in the ground-truth segmentation as text queries.

queries, *e.g.*, 847 classes when evaluated for all images in `A-847`. The Grounding mIoU evaluates concept grounding. An example scenario is interactive segmentation where users can specify a set of concepts in an image for the model to segment. It only uses the ground-truth classes in an image, *e.g.*, 7 classes are used as text queries for the example in the second row of Figure 5. We find that predictions in the mIoU and Grounding mIoU settings can look quite differently and sometimes mIoU does not correctly reflect the prediction quality due to class ambiguity. For example, building, brick, house are all correct visual concepts to describe the object in Figure 5 but the ground-truth label is building.

Zero-shot transfer to ADE20k/PASCAL: We evaluate the performance of OpenSeg and the baselines on holdout image segmentation datasets whose train sets are not used for training. In Figure 5 (bottom), we compare ALIGN, ALIGN w/proposal, LSeg+ and OpenSeg on the challenging `A-847` and `PC-459` datasets with large vocabularies and also on the widely used `A-150` and `PC-59`. In the following sections we discuss our findings based on these results.

OpenSeg significantly outperforms pre-trained ALIGN [23]: OpenSeg trained on COCO outperforms ALIGN baseline on all of the benchmarks significantly. While adding proposals to ALIGN improves mIoU results. OpenSeg still performs significantly better. For example, on `PC-459` OpenSeg outperforms ALIGN and ALIGN w/proposals by +5.4 and +4.2 mIoU, respectively.

Training on limited categories hurts generalization: LSeg+, which is trained with pixel-wise segmentation in COCO, outperforms ALIGN by a large margin on COCO (+39.5 mIOU) and `PC-59` (+28.0 mIOU). Note COCO categories contain most of `PC-59` categories. However, when we evaluate LSeg+ on

A-847 which includes a larger set of vocabularies, the performance of LSeg+ is worse than ALIGN by 1.0 mIoU and 7.3 Grounding mIoU. These results demonstrate that training on the limited categories of COCO hurts the generalization of the model.

OpenSeg improves generalization: . While OpenSeg trained on COCO has worse mIoU on COCO and PC-59 in comparison to LSeg+, it generalizes better on all other benchmarks. OpenSeg outperforms LSeg+ by +2.5 mIoU and +11.3 Grounding mIoU on A-847 and also by +1.2 mIoU and +15.0 Grounding mIoU on PC-459. The OpenSeg uses class-agnostic masks and image-level caption supervision, while LSeg+ uses 134 per-pixel class name supervision. Although OpenSeg is trained with a weaker supervision, it has a better generalization to classes outside of COCO. These results reveal that we need open-vocabulary supervision such as captions for training a *generalist* model.

Scaling training data with captions improves performance: To scale up training data we utilize the Localized Narrative dataset, which includes detailed narratives about the objects and stuff in each image. We train a segmentation teacher model on the COCO dataset and use it to generate segmentation pseudo labels on the Loc. Narr. dataset. By scaling training data from 118k images to 652k images, the performance of OpenSeg improves on average by 2.5 mIoU and 4.8 Grounding mIoU across 4 benchmarks (see Figure 5). In Appendix G, we study the importance of using pseudo segmentation labels during training.

Ensembling of text queries and prompt engineering: To further improve the performance of OpenSeg we use ensembling where we include synonyms or subcategories of classes. For example, we use 'person', 'child', 'girl', 'boy', *etc.* for the class of 'person'. We ensemble the multiple text queries by taking the max score as described in the Section 3.4. Also, since some of the class names of the segmentation datasets are not descriptive, we add a short context to the names. *e.g.* we change 'glass' to 'drinking glass'. These improvements give us on average 2.6 mIoU gain across 4 datasets (see Table 2). See Appendix I for more details.

Compare with existing methods: We compare OpenSeg with previous open-vocabulary and zero-shot segmentation methods in Table 2. We initialize ResNet101 backbone of OpenSeg and LSeg+ with ImageNet pretrained weights similar to the baselines. LSeg+ significantly outperforms LSeg (and also SPNet [49] and ZS3Net [6]) as it is trained on the larger dataset of COCO instead of PASCAL-20. In contrast to LSeg and LSeg+ which are trained on COCO class labels, OpenSeg is trained on COCO captions and as a result has a better generalization. OpenSeg outperforms LSeg+ by +1.3 mIoU on PC-459. Compared with GroupVit, OpenSeg learns visual grouping with class-agnostic segmentation, and has a superior performance. Also, by scaling up the training data from COCO to COCO+Loc. Narr. it achieves further gain of +1.4 on PC-459.

For the strongest OpenSeg (last two rows), we initialize EfficientNet-b7 backbone with ALIGN pre-trained image encoder [23]. Also we train this model with sync loss (see Appendix F for more details). This model significantly outperforms the strongest LSeg model with ViT-L backbone (+19.9 mIoU on PASCAL-20).

Table 2. The mIoU results of our model and previous open-vocabulary and zero-shot segmentation methods. Results for SPNet and ZS3Net on PASCAL-20 are reported from [29].

	backbone	external dataset	target dataset	A-847	PC-459	A-150	PC-59	PAS-20
LSeg [29]	ViT-L/16	✗	✓(seen classes)	-	-	-	-	52.3
SPNet [49]	ResNet101	✗	✓(seen classes)	-	-	-	24.3	18.3
ZS3Net [6]	ResNet101	✗	✓(seen classes)	-	-	-	19.4	38.3
LSeg [29]	ResNet101	✗	✓(seen classes)	-	-	-	-	47.4
LSeg+	ResNet101	COCO	✗	2.5	5.2	13.0	36.0	59.0
OpenSeg(ours)	ResNet101	COCO	✗	4.0	6.5	15.3	36.9	60.0
OpenSeg(ours)	ResNet101	COCO+Loc. Narr.	✗	4.4	7.9	17.5	40.1	63.8
GroupVit [51]	VIT-S	CC12M+YFCC	✗	-	-	-	22.4	52.3
OpenSeg(ours)	eff-b7	COCO+Loc. Narr.	✗	8.1	11.5	26.4	44.8	70.2
+prompt eng.	eff-b7	COCO+Loc. Narr.	✗	**8.8**	**12.2**	**28.6**	**48.2**	**72.2**

4.4 Ablation Experiments

Importance of backbone initialization: In order to save the computation, we initialize OpenSeg from the state-of-the-art ALIGN checkpoint trained on 1.8 billion examples for image-text alignments. In this section, we study the importance of initialization of the vision backbone from this checkpoint. In Table 3, we compare the performance of training OpenSeg from scratch, initializing from the NoisyStudent checkpoint [50] and initializing from the ALIGN checkpoint. For training these models, we use the same hyper-parameters, and only tune the learning rate (0.32 for scratch, 0.08 for NoisyStudent init. and 0.005 for ALIGN init.) and number of steps (180k steps for scratch and 60k for NoisyStudent and ALIGN init.). Table 3 shows that using the NoisyStudent checkpoint to initialize the backbone achieves slightly worse results (less than 0.5 mIoU on all benchmarks) compared to using the ALIGN checkpoint. This shows initializing from the ALIGN model is not necessary for good word-region alignments. However, training from scratch is still trailing behind. We may be able to reduce the gap by increasing the batch size and training with more data.

Table 3. Backbone initialization with an ALIGN pre-trained image encoder is not critical. The models use the pre-trained ALIGN text encoder and are trained on COCO and Loc. Narr. datasets.

	A-847	PC-459	A-150	PC-59
Random init.	4.5	7.6	18.6	40.6
NoisyStudent init.	6.6	10.7	24.4	**46.9**
ALIGN init.	**6.8**	**11.2**	**24.8**	45.9

Incorporating proposals at inference time improves accuracy: We are curious about the importance of mask proposals in OpenSeg during inference. To study this problem, we take the feature map $\mathcal{F}_z$ in OpenSeg and perform per-pixel segmentation by taking the dot product of $\mathcal{F}_z$ with word embeddings $\mathbf{w}$. This method performs inference without mask proposals. In Table 4, we compare the performance of OpenSeg and its counterparts that do not use mask proposals (the above method) or using ground-truth as mask proposals. The performance

Table 4. Incorporating predicted masks at inference improves mIoU accuracy. Using the ground-truth masks can be seen as the performance upper bound when segmentation masks are perfectly predicted. The model is trained on COCO.

	A-847	PC-459	A-150	PC-59
OpenSeg	6.3	9.0	21.1	42.1
- pred. masks	(-1.7) 4.6	(-3.1) 5.9	(-4.7) 16.4	(-10.0) 32.1
+ gt. masks	(+2.8) 9.1	(+3.3) 12.3	(+6.4) 27.5	(+7.2) 49.3

Table 5. Using all words in training captions hurts performance. Using nouns+adj for training achieves the best results. The model is trained on COCO.

caption filter	A-847	PC-459	A-150	PC-59
all words	5.3	8.8	20.0	41.3
noun + adj. + verb	6.0	8.8	20.9	41.7
noun + adj	**6.3**	**9.0**	**21.1**	**42.1**

of OpenSeg is much worse if not using proposals: mIoU on `PC-59` drops from 42.1 to 32.1 and from 21.1 to 16.4 on `A-150`. Using ground-truth as proposals can be seen as an upper bound when we have perfect class-agnostic localization. The results show the room for improving localization. It also demonstrates even with perfect localization, the semantic alignment is still challenging.

Importance of text filtering: We train OpenSeg with image captions which may include words that do not represent any regions in an image. These noises make training more challenging. We perform a simple pre-processing on the captions and extract the list of nouns and adjectives. This procedure removes conjunctions, pronouns, adverbs, verbs, *etc.* which reduces the noises. In Table 5, we study the performance of OpenSeg when using different types of filtering on the captions. Keeping only nouns and adjectives yields the best results. The worst results are from using all words, which show 0.2-1.1 worse mIoU. The small performance differences across different ways of text filtering show OpenSeg is robust to the noise in the input words to some degree.

5 Conclusion

We propose OpenSeg, an open-vocabulary image segmentation model, to organize an image into regions described with arbitrary text queries. This is in stark contrast to previous works in semantic segmentation learned to predict categories in closed vocabulary. We propose to represent an image with a set of mask regions followed by visual-semantic alignments. Such representations support weakly-supervised learning for grounding words in a caption to predicted mask proposals, and thus make the training data scalable. We are the first work to directly evaluate on holdout image segmentation datasets, attaining significant performance gains against strong baselines initialized by a pre-trained ALIGN model. We hope to encourage future works to learn a *generalist* segmentation model that can transfer across datasets using language as the interface.

References

1. Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J.W., Brundage, M.: Evaluating clip: towards characterization of broader capabilities and downstream implications. arXiv preprint arXiv:2108.02818 (2021) 18
2. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. PAMI (2010) 4
3. Arbelaez, P., Pont-Tuset, J., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: CVPR (2014) 3, 4, 9
4. Baek, D., Oh, Y., Ham, B.: Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In: ICCV (2021) 4
5. Bolei Zhou, Hang Zhao, X.P.T.X.S.F.A.B., Torralba, A.: Semantic understanding of scenes through ade20k dataset. IJCV (2018) 4
6. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. NeurIPS (2019) 4, 12, 13
7. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018) 4
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV (2020) 3
9. 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) 5, 8
10. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. arXiv (2021) 3, 6, 9, 18
11. Comaniciu, D., Meer, P.: Robust analysis of feature spaces: Color image segmentation. In: CVPR (1997) 4
12. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: ICCV (2021) 5
13. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010) 4, 9, 10
14. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: CVPR (2015) 5
15. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016) 5
16. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR (2021) 8
17. Ghiasi, G., Zoph, B., Cubuk, E.D., Le, Q.V., Lin, T.Y.: Multi-task self-training for learning general representations. In: ICCV (2021) 6, 7
18. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Zero-shot detection via vision and language knowledge distillation. arXiv e-prints pp. arXiv–2104 (2021) 3
19. Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D.: Contrastive learning for weakly supervised phrase grounding. In: ECCV (2020) 5, 7
20. Hu, P., Sclaroff, S., Saenko, K.: Uncertainty-aware learning for zero-shot semantic segmentation. In: NeurIPS (2020) 4
21. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: ECCV (2016) 5

22. Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H.: Bi-directional relationship inferring network for referring image segmentation. In: CVPR (2020) 5
23. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. ICML (2021) 2, 3, 5, 8, 10, 11, 12
24. Jo, S., Yu, I.J.: Puzzle-cam: Improved localization via matching partial and full features. arXiv preprint arXiv:2101.11253 (2021) 4
25. Kamath, A., Singh, M., LeCun, Y., Misra, I., Synnaeve, G., Carion, N.: Mdetr–modulated detection for end-to-end multi-modal understanding. arXiv preprint arXiv:2104.12763 (2021) 5
26. Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P.: Panoptic segmentation. In: CVPR (2019) 3, 8
27. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017) 5
28. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: MSeg: A composite dataset for multi-domain semantic segmentation. In: CVPR (2020) 4, 9, 19
29. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. ICLR (2022) 2, 4, 5, 10, 13
30. Li, P., Wei, Y., Yang, Y.: Consistent structural relation learning for zero-shot segmentation. NeurIPS (2020) 4
31. Li, Y., Kuang, Z., Liu, L., Chen, Y., Zhang, W.: Pseudo-mask matters in weakly-supervised semantic segmentation. In: ICCV (2021) 4
32. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017) 6, 8
33. Maninis, K.K., Pont-Tuset, J., Arbeláez, P., Gool, L.V.: Convolutional oriented boundaries: From image segmentation to high-level tasks. TPAMI (2018) 3, 4, 9
34. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV (2016) 6
35. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR (2014) 4, 9
36. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR (2015) 4
37. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015) 5
38. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: ECCV (2020) 9
39. Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Lin, Z., Torr, P., Jia, J.: Open-world entity segmentation. arXiv preprint arXiv:2107.14228 (2021) 4
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. ICML (2021) 2, 3, 4, 5, 10
41. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: ECCV (2016) 5
42. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI (2000) 4
43. Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation. In: Proceedings of the 5th Conference on Robot Learning (CoRL) (2021) 18

44. Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: ICML (2019) 8
45. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. IJCV (2013) 4
46. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: MaX-DeepLab: End-to-end panoptic segmentation with mask transformers. In: CVPR (2021) 3, 6
47. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: CVPR (2020) 4
48. Wertheimer, M.: Laws of organization in perceptual forms. In: Ellis, W. (ed.) A Source Book of Gestalt Psychology, pp. 71–88. Routledge and Kegan Paul, London (1938) 2, 4
49. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: CVPR (2019) 4, 12, 13
50. Xie, Q., Hovy, E., Luong, M.T., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR (2020) 13
51. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. CVPR (2022) 5, 13
52. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Sohel, F., Xu, D.: Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In: ICCV (2021) 4
53. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. arXiv preprint arXiv:2112.14757 (2021) 5
54. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: CVPR (2019) 5
55. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR (2018) 5
56. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016) 5
57. Zabari, N., Hoshen, Y.: Semantic segmentation in-the-wild without seeing any segmentation examples. arXiv preprint arXiv:2112.03185 (2021) 5
58. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR (2021) 3, 7
59. Zhao, H., Puig, X., Zhou, B., Fidler, S., Torralba, A.: Open vocabulary scene parsing. In: ICCV (2017) 4
60. Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. CVPR (2016) 2, 4, 10
61. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV (2019) 9
62. Zhou, C., Loy, C.C., Dai, B.: Denseclip: Extract free dense labels from clip. arXiv preprint arXiv:2112.01071 (2021) 5