Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Language Models

Jiayun Luo[♭], Siddhesh Khandelwal[♯], Leonid Sigal[♯], and Boyang Li[♭]
[♭]Nanyang Technological University, Singapore
[♯]University of British Columbia, Vector Institute for AI, Canada
{luoj0028, boyang.li}@ntu.edu.sg, {skhandel, lsigal}@cs.ubc.ca

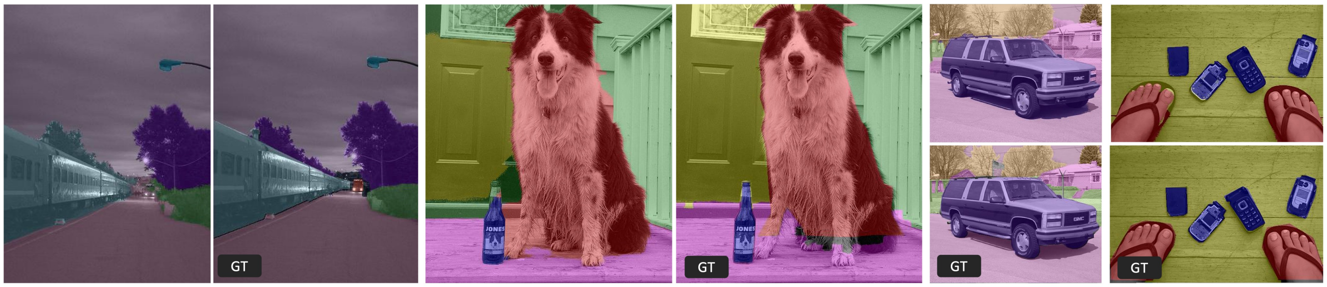


Figure 1. **Qualitative Results of PnP-OVSS + BLIP.** Images are from Pascal Context and COCO stuff. The right columns and bottom rows show the ground-truth (GT); the rest are our results. Note accurate results even on complex (trees) and small objects (last column).

Abstract

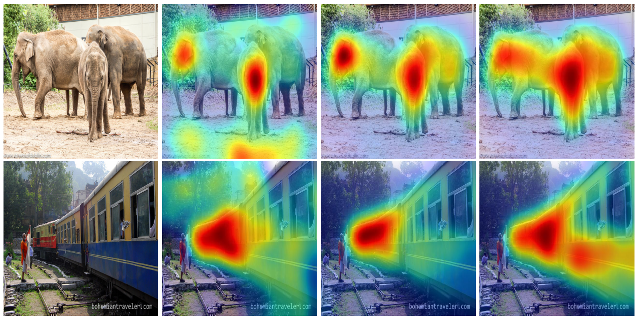
From image-text pairs, large-scale vision-language models (VLMs) learn to implicitly associate image regions with words, which prove effective for tasks like visual question answering. However, leveraging the learned association for open-vocabulary semantic segmentation remains a challenge. In this paper, we propose a simple, yet extremely effective, training-free technique, Plug-and-Play Open-Vocabulary Semantic Segmentation (PnP-OVSS) for this task. PnP-OVSS leverages a VLM with direct text-to-image cross-attention and an image-text matching loss. To balance between over-segmentation and under-segmentation, we introduce Salience Dropout; by iteratively dropping patches that the model is most attentive to, we are able to better resolve the entire extent of the segmentation mask. PnP-OVSS does not require any neural network training and performs hyperparameter tuning without the need for any segmentation annotations, even for a validation set. PnP-OVSS demonstrates substantial improvements over comparable baselines (+29.4% mIoU on Pascal VOC, +13.2% mIoU on Pascal Context, +14.0% mIoU on MS COCO, +2.4% mIoU on COCO Stuff) and even outperforms most baselines that conduct additional network training on top of pretrained VLMs. Our codebase is at https://github.com/letitiabanana/PnP-OVSS.

1. Introduction

The classic task of semantic segmentation [20, 22] aims to classify pixels to their object types. Traditional supervised methods require dense pixel-level annotations and are restricted to recognizing a predefined set of objects. To relax these constraints, open-vocabulary semantic segmentation [8, 17, 21, 41, 44, 46, 48, 75, 77, 78, 81, 85] aspires to identify arbitrary object categories, whereas weakly supervised techniques [9, 11, 15, 25, 31, 36, 53, 55, 66, 76, 91] can acquire pixel-level localization capabilities from coarse supervision, *e.g.*, image labels or boxes.

Large-scale vision-language models (VLMs) pretrained on image-text pairs [1, 37, 38, 40, 61, 79, 86] achieve unprecedented performance on multimodal tasks, such as describing arbitrary images and answering free-form, open-ended questions about them (either with [37, 38, 61] or without finetuning [1, 19, 63, 71]). These tasks apparently involve some ability to localize objects. For example, to answer the question "*what objects appear on the table?*", the model would have to first localize the table in the image and identify the objects on it. Hence, it is reasonable to conjecture that the VLM network learns to perform open-vocabulary localization from image-text pretraining. However, distilling the localization capability from the VLMs remains an open challenge.

Most existing methods for open-vocabulary semantic



(a) Original image (b) Cross-attention (c) GradCAM (d) Salience DropOut

Figure 2. Segmentation masks for *elephant* and *train* using (a) off-the-shelf cross-attention, (b) cross-attention + GradCAM, and (c) cross-attention + GradCam + Salience DropOut (§3.2). The naive cross-attention masks are too inclusive whereas GradCAM is too exclusive.

segmentation (OVSS) from VLMs usually obtain single vector encodings for the visual and text inputs respectively [43, 44, 51, 52, 74, 75]. However, pooling every token into a single vector likely discards information about detailed positions of objects and words. We investigate the use of pretrained cross-attention layers for OVSS, which retain finer-grained correspondence between text and image patches.

Nevertheless, a naive application of cross-attention between the class name and the image patches lead to overly broad segmentation masks that include irrelevant parts of the image (*i.e.*, over-segmentation, see Fig. 2 (b)). To alleviate this, [63] employs gradient information from the image-text matching loss, GradCAM[57]-style, to sharpen the attention-based masks for the purpose of guiding caption generation. However, this results in masks that capture only the most discriminative regions of an object, such as the head of an elephant (*i.e.*, under-segmentation, Fig. 2 (c)). To acquire complete object masks, we propose Salience DropOut, which iteratively drops the image patches with high GradCAM attention scores, forcing the model to attend to less discriminative but relevant object parts.

Another important consideration is the cross-attention layer and the attention head to extract the masks from. These hyperparameters have enormous influence on the final results and are traditionally tuned on a validation set with pixel-level mask annotations. To eliminate the need for dense annotations, we propose an weakly-supervised reward function based on CLIP [49]. On a validation set with images as well as object class names, the technique contrasts the extracted object regions with a blank image. If, according to CLIP, the former is more similar to the corresponding class name than the latter, we increment the reward. All hyperparameter tuning of our technique is performed with a simple random search with this reward, leading to high performance.

In summary, we propose Plug-and-Play Open-vocabulary Semantic Segmentation (PnP-OVSS), an extremely simple and training-free framework to extract semantic segmentations from VLMs. At zero extra training cost, PnP-OVSS can be used with any pretrained VLM with text-to-image cross attention layer and an image-text matching loss. It has zero reliance on pixel-level annotations, including a validation set for hyperparameter tuning. At the same time, PnP-OVSS delivers excellent performance. It not only beats the training-free baseline with remarkable margins (+29.4% mIoU on Pascal VOC, +13.2% on Pascal Context, +14.0% on MS COCO, +2.4% on COCO Stuff), but also outperforms most recent techniques within the past two years that require extensive finetuning on top of the VLM pretraining.

With this paper, we make three contributions:

- We propose to combine text-to-image attention, GradCAM, and Salience DropOut to iteratively acquire accurate segmentation of arbitrary classes from pretrained VLM.
- We replace the densely annotated validation set for hyperparameter tuning, which is needed by most existing methods, with a contrastive reward function based on CLIP. This reward function, coupled with random search, finds a good set of hyperparameters for OVSS.
- The proposed method, PnP-OVSS, is simple to use, requires no extra finetuning, and delivers high performance. Its success hints at a new direction for open-vocabulary segmentation tasks leveraging large VLMs.

2. Related Work

2.1. Large-Scale Vision-Language Model

Large-scale vision-language models (VLMs), trained on millions of image-text pairs, have become the foundation for many multimodal tasks. Architecturally, the straightforward approach to training such methods involves aligning visual and textual latent representations via a simple dot product [24, 49]. However, this is insufficient for complex structured tasks like visual question answering or image captioning, which require specialized approaches that employ separate encoders before cross-attention between modalities [33, 35, 37, 79, 80] or self-attention networks over all tokens from both modalities [10, 27, 39, 40, 61]. Another design dimension when training VLMs is the loss function. Commonly used losses include image-text contrastive learning [34, 35, 37, 50, 61, 82], image-text matching (ITM) [27, 34, 35, 37, 61, 73, 79, 82, 83], prediction of masked tokens or patches [34, 61, 79, 82], and language modeling [1, 7, 34, 37].

This work utilizes models with unimodal encoders followed by cross-attention fusion [34, 35, 37, 73, 79, 82, 83], as they work with high-level features, and accurately attend to the appropriate image patches. Additionally, we utilize the gradient from the ITM loss in the GradCAM step (§3.1)

Method	PT Data Size	FT data size
Requires finetuning on image-text pairs		
OVSegmentor [75]	-	4.3M
Vil-Seg [43]	400M	412M
GroupVit* [74]	-	3.4M
GroupVit [74]	-	26M
CLIPpy [51]	-	134M
SegClip [44]	400M	3.4M
ViewCo [52]	-	26M
TCL [6]	400M	15M
PACL [46]	400M	30M
Requires finetuning but not image-text pairs		
MaskClip w/ ST [90]	400M	1.2M
ZeroSeg* [8]	400M	3.4M
ZeroSeg [8]	400M	1.2M
Requires no finetuning		
MaskClip[90]	400M	0
Reco[58]	400M	0
PnP-OVSS (Ours)		
+ BLIP	129M	0
+ BridgeTower	400M+ 4M	0

Table 1. Training data of current Zero-shot semantic segmentation methods with only text supervision. PT stands for pretraining with image-caption data, FT stands for fine-tuning with image-caption data. ST stands for self training. We list only the image-caption data used for pretraining and all type of data for finetuning. For hyperparameter tuning, our method uses CLIP-L/14 pretrained with 400M data to calculate the reward.

to sharpen segmentation masks.

2.2. Zero-shot Semantic Segmentation

Zero-shot semantic segmentation predicts a dense segmentation mask for any object class described by a given text prompt, with only prior exposure to class-agnostic image-level supervision. This contrasts with weakly supervised semantic segmentation [3, 13, 23, 26, 28, 30, 54, 62, 70, 70, 84, 87, 88], which relies on class-specific annotations, and unsupervised object discovery [12, 14, 56, 59, 64, 65, 68, 69], which identifies the sole object in the foreground.

Traditional methods [4, 18, 47, 72] train a classifier to distinguish between seen and unseen visual features, wherein the unseen visual features are obtained from a generative model, trained on pairs of seen class image-text embeddings. Recently, methods additionally leverage knowledge from VLMs to attain better matching of visual and textual features [17, 32, 41, 77, 90]. Concretely, they train the segmentation network with dense annotations, while replacing a part of the framework with components from VLMs.

Recently, methods have explored the use of additional supervision to further reduce the need for pixel level annotations. Prior work is compared in Table 1.

Models Finetuned on Image-Text Pairs. Methods proposed in [6, 43, 44, 46, 51, 52, 74, 75] require paired image-text annotations to adapt to the zero-shot segmentation task. Specifically, approaches in [43, 44, 51, 52, 74, 75] cluster semantically similar pixels by contrasting grouped region embeddings with textual embeddings, usually via the use of contrastive and self-supervised losses. PACL [46] modifies the constrastive loss to operate over aggregated patch embeddings (instead of image embeddings) to encourage better alignment between image patches and text. TCL [6] achieves a similar patch-text alignment by introducing an additional module to extract text-grounded image regions.

Models Finetuned on Pseudo-labels. Methods in [8, 90] do not require additional image-text supervision, but involve additional fine-tuning to adapt to the segmentation task. ZeroSeg [8] attempts to match its own group embeddings with multi-scaled segment embeddings obtained from CLIP [49]. MaskClip w/ ST [90] modifies the global attention pooling layer within CLIP to output segmentation masks, which are used as pseudo-labels to train a segmentation network.

In comparison, our proposed PnP-OVSS does not require *any fine-tuning or additional paired image-text annotations*. It can directly distill high quality open vocabulary semantic segmentations from any VLM with direct text-to-image cross-attention and an image-text matching loss [34, 35, 37, 73, 79, 82, 83]. Compared to approaches in [58, 90] that perform zero-shot open vocabulary semantic segmentation under a similar no training and no additional annotations paradigm, PnP-OVSS achieves considerably superior performance.

3. Method

PnP-OVSS has four major steps. First, we extract a cross-attention salience map per object class from a VLM. Second, we sharpen the salience map by weighing it with the ITM gradient in the style of GradCAM. Third, we apply Salience DropOut that iteratively completes the salience maps. Fourth, we apply Dense CRF [29] for fine-grained adjustment. The process is illustrated in Fig. 3 and Fig. 4.

In the first step, we feed an image and a text prompt to the pretrained VLM and extract cross-attention maps. The text prompt is “A picture of `[class 1] [class 2]` ... `[class K]`”, which includes all K class names of interest in the dataset. The image is divided into $P \times P$ patches. The text prompt and the image first go through the modality-specific encoders respectively, followed by a cross-attention fusion module. In the cross-attention layers, the text encodings serve as query vectors and the image patche encodings are as key and value vectors.

We extract an attention map for each text token to the image patches from a particular cross-attention layer and an

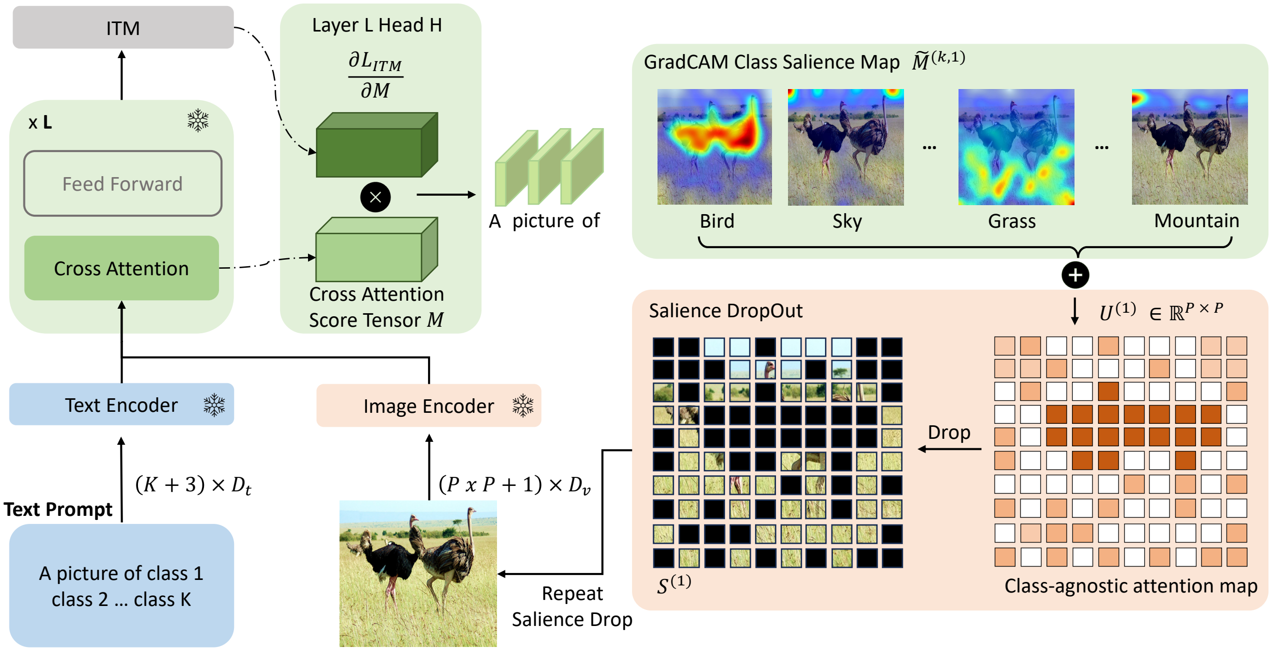

ITM
Layer L Head H
$\frac{\partial L_{ITM}}{\partial M}$
x L
Feed Forward
Cross Attention
Cross Attention Score Tensor M
A picture of
GradCAM Class Salience Map $\widetilde{M}^{(k,1)}$
Bird
Sky
...
Grass
...
Mountain
Salience DropOut
$U^{(1)} \in \mathbb{R}^{P \times P}$
Drop
$S^{(1)}$
Class-agnostic attention map
Text Encoder
Image Encoder
$(K+3) \times D_t$
$(P\ x\ P+1) \times D_v$
Text Prompt
A picture of class 1 class 2 ... class K
Repeat Salience Drop

Figure 3. The first iteration with cross-attention + GradCAM + Salience DropOut. The text prompt contains K class names and the image contains $P \times P$ patches. From a cross-attention layer and an attention head in the pretrained VLM, we obtain K attention score maps of size $P \times P$, which are sharpened by GradCAM using gradients from the image-text-matching (ITM) loss. To get more complete predictions, we perform Salience Dropout, which repeatedly zero out image patches of the highest average scores and feeds the remaining patches to the image encoder again, forcing the model to attend to other less discriminative patches. We show example salience maps from all iterations in Fig. 4.

attention head. Note that different layers and heads lead to drastic performance differences, and the choices are hyperparameters, tuned using the procedure in §3.4. We exclude attention maps for the first three tokens “A”, “picture”, “of”, which do not describe semantic classes. If a class name contains two or more tokens, we take the mean attention map. This procedure yields an attention tensor of size $K \times P \times P$.

With the correct layer and attention head, we observe that this attention map, when normalized by `softmax` along the first dimension K, can provide passable semantic segmentations. However, it tends to include many patches unrelated to the class being segmented, leading to over-segmentation. Hence, we introduce two refinement steps, GradCAM and Salience DropOut, explained in §3.1 and §3.2. After these steps, we acquire aggregate salience maps for every class. Finally, we conduct local polish to the resultant salience maps using Dense CRF, as described in §3.3.

3.1. Map Sharpening with GradCAM

The off-the-shelf attention maps tend to also cover many patches unrelated to the class name (See Fig. 2 (b)). Prior works [35, 63] further sharpen the attention maps and focus them on class-discriminative regions using a variant of GradCAM [57], which is originally proposed for convolutional networks, but applicable to attention maps. It is worth noting that [35, 63] use the technique for purposes other than semantic segmentation.

The GradCAM method requires a gradient. Here we leverage the image-text matching (ITM) loss, which trains the VLM to classify if an image-text pair match each other or not. Computing the ITM loss requires a label. We use “matching” (as opposed to “not matching”) as the label and compute the gradient of the loss with respect to the attention score. This is equivalent to asking: *which attention scores contribute the most to the decision that the image-text pair is matching?* Formally, we denote a $P \times P$ attention map for class k as $M^{(k)}$ and the ITM loss as $\mathcal{L}_{\text{ITM}}$. The GradCAM class salience map is computed as

$$\widetilde{M}^{(k)} = \max\left(0, \frac{\partial \mathcal{L}_{\text{ITM}}}{\partial M^{(k)}}\right) \otimes M^{(k)}, \tag{1}$$

where $\otimes$ denote the component-wise multiplication and $\max(\cdot)$ is also applied component-wise.

3.2. Salience DropOut

As illustrated in Fig. 2 (c), segmentations generated by the GradCAM-style re-weighting of cross-attention are often narrowly focused on the most discriminative regions for a given class. However, the less discriminative regions are

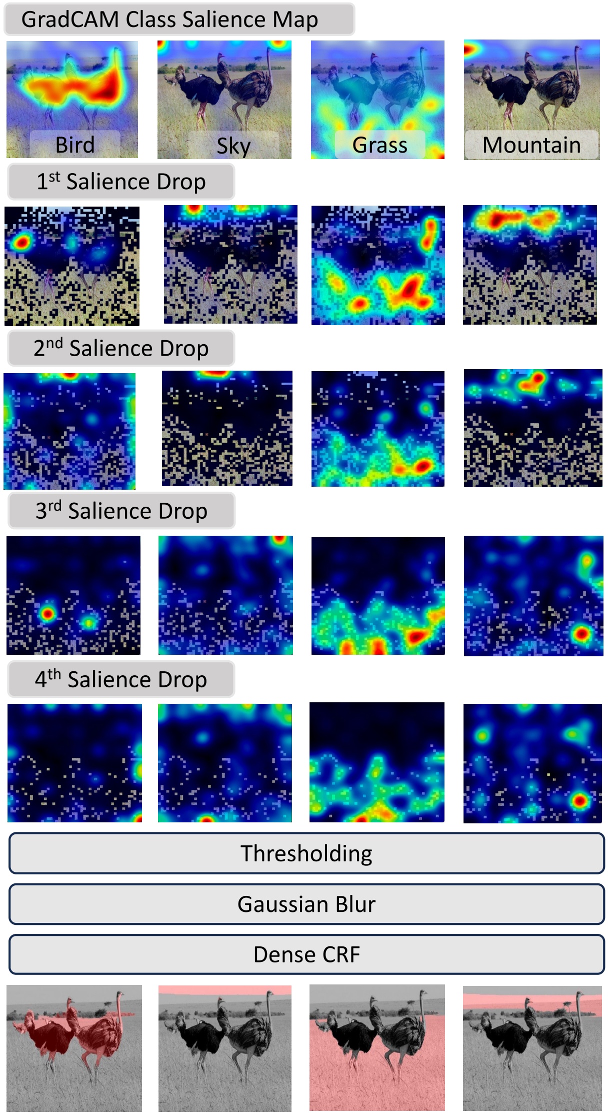


Figure 4. An illustration of Salience DropOut, showing GradCAM salience values after each iteration. Black squares in the images indicate dropped patches. We obtain the final result by summing the salience maps from all iterations and applying thresholding, Gaussian blur, and Dense CRF.

still important for the completeness of masks. To compel the VLM to attend to these regions, we propose an iterative technique called Salience DropOut.

Since our task is zero-shot and open-vocabulary, we do not have prior knowledge of the classes present in the image. Hence, we sum up the salience maps $\widetilde{M}^{(k)}$ over all classes k, yielding a class-agnostic salience map $U^{(t)}$ for the t^{th} Salience DropOut iteration, $U^{(t)} = \sum_{k=1}^{K} \widetilde{M}^{(k,t)}$, where $\widetilde{M}^{(k,t)}$ is the GradCAM salience map for class k after the t^{th} iteration. Next, we zero out the 50% of image patches with the highest values in $U^{(t)}$. On the remaining image, we compute the GradCAM salience maps $\widetilde{M}^{(k,t+1)}$ and their sum, $U^{(t+1)}$. Any image patch previously zeroed out will always receive zero salience in later iterations.

Formally, the set of remaining image patches $S^{(t)} \subseteq \{1, \dots, P\}^2$ after the t^{th} dropout iteration is defined as

$$S^{(t)} = S^{(t-1)} \setminus \{(i,j) \mid U_{ij}^{(t)} \geq \eta\}, \tag{2}$$

$$\eta = \texttt{median}\left(\{U_{ij}^{(t)} \mid (i,j) \in S^{(t-1)}\}\right), \tag{3}$$

where $U_{i,j}^{(t)}$ denotes the aggregate salience value of the image patch at row i and column j. Additionally, note that the input to the first dropout iteration $S^0 = \{1, \dots, P\}^2$ is the set of all image patches.

We stop at four rounds of dropout, as almost all (93.75%) patches are removed beyond that point. The final output for each class k is the sum over all salience maps across the four dropout iterations, $\hat{M}^{(k)} = \sum_{t=1}^{4} \widetilde{M}^{(k,t)}$.

3.3. Gaussian Blur and Dense CRF

The salience dropout procedure generates k continuous-valued salience maps, one per object class. To filter out small random noise in the salience values, we subsequent apply a straightforward thresholding operation at a predefined value T on the salience values and obtain binary segmentation masks. Nonetheless, the hard thresholding creates jagged segmentations with sharp edges that often do not coincide with object boundaries. One common strategy in zero-shot segmentation is to apply Dense Conditional Random Field (CRF) [29], which makes fine-grained adjustments to the estimated masks by enforcing consistency between nearby image pixels with similar colors.

However, we find that the hard 0/1 labels in the binary masks do not work well as pixel unary potentials for Dense CRF. Hence, we smooth them using a Gaussian kernel with a preset variance σ, which results in a better initialization for unary terms and accounts for uncertainty of exact segmentation boundary along the patch boundaries.

3.4. Hyperparameter Tuning

Three hyperparameters in PnP-OVSS have the strongest influence on the result, the cross-attention layer L, the attention head H, and the binary threshold T. Traditionally, tuning these hyperparameters requires a validation set with pixel-level labels. However, since our goal is to perform zero-shot open-vocabulary semantic segmentation, this requirement could potentially limit the applicability of the technique. Instead, we propose a weakly supervised reward function for hyperparameter tuning, which only requires a set of images and the class names appearing in each image.

The reward for an image I is calculated as follows. We start with a set of classes present in the image, denoted as $\mathcal{K}(I)$. For each class $k \in \mathcal{K}(I)$, we obtain a segmentation mask $M^{(k)}$ (which can be the GradCAM mask, the Salience DropOut mask, or the Dense CRF mask). Next, we apply the mask to the image I and input the extracted regions $M^{(k)} \otimes I$ into a pretrained neural network f, which takes an image and a textual class name as input and produces a similarity score. We calculate the normalized probability that the masked image $M^{(k)} \otimes I$ belongs to the ground-truth class k and contrast with a completely black image $\mathbf{0}$.

$$\text{Reward} = \sum_{k \in \mathcal{K}(I)} \mathbb{1}[Pr(M^{(k)} \otimes I, k) > Pr(\mathbf{0}, k)], \quad (4)$$

$$Pr(I, k) = \frac{\exp(f(I, k))}{\sum_{k' \in \mathcal{K}(I)} \exp(f(I, k'))}, \quad (5)$$

where $\mathbb{1}(\cdot)$ is the indicator function. Intuitively, a reward of 1 is assigned if and only if the image features pulled with the estimated mask for class k bears higher similarity to the class name of k than a black image (which can be interpreted as the prior probability of class k).

We sum up the reward for all validation images as the total reward. The best hyperparameters, including the cross-attention layer, the attention head, the threshold T, and the variance in the Gausian blur kernel, are determined using a simple random search.

4. Experiments

4.1. Datasets and Implementation Details

Following the previous work for zero-shot semantic segmentation, we adopt validation sets of Pascal-VOC 2012 [16], Pascal Context [45], COCO Object [42], COCO Stuff [5], and ADE20K [89] that contain, respectively, 20 object classes, 59 object and stuff classes, 80 object classes, 171 object and stuff classes, and 150 object and stuff classes to evaluate our framework. To verify its versatility, we apply PnP-OVSSto two high-performance VLMs, BLIP [37] and BridgeTower [79], which have the ITM loss and text-to-image cross-attention. More details are in the supplementary material.

Hyperparameter tuning. We use CLIP VIT-L/14 to calculate the reward. The input resolution is 336×336. The search spaces of the random search and results are shown in Tab. 2. For computation efficiency, we tune the layer, the head, and the threshold with GradCAM masks. With the first three hyperparameters fixed, we tune the Gaussian variance on masks before Dense CRF. We directly adopt Dense CRF hyperparameters from CutLER [67] without tuning.

Baselines. We adopt recent papers on zero-shot open-vocabulary semantic segmentation as baselines. The only baselines strictly comparable to our work are MaskClip [90] and Reco [58], which do not perform any finetuning on pretrained VLMs and do not perform hyperparameter tuning on dense annotations. We call these Group 3. To further expand our scope, we also include two other groups of baselines. Group 2 finetunes VLMs but does not require image-text pairs, including MaskClip with self-training (ST) [90] and ZeroSeg [8]. We include ZeroSeg, trained with ImageNet1K, and ZeroSeg*, trained with CC3M+COCO. Group 1 contains baselines that require training on image-text pairs. For completeness, we also compare against supervised techniques since 2019. To maintain the zero-shot setting, we test them on classes not observed during training. For more details, we refer readers to the the supplementary material.

Hyperparameters	Start	End	Step	Solution
BLIP				
Layer	1	12	1	8
Head	1	12	1	10
Attention Threshold	0.05	0.5	0.1	0.15
BridgeTower				
Layer	1	6	1	2
Head	1	16	1	8
Attention Threshold	0.05	0.5	0.1	0.15
Gaussian Blur				
Standard Deviation	0.01	0.11	0.02	0.05

Table 2. Search space for hyperparameters

4.2. Main Results

We show the main results in Tab. 3. As input resolution may influence results and cause unfair comparisons, we label the resolution used by each method. PACL [46] uses an 224 resolution but changes the stride for image patchification from 16 to 4, hence introducing overlapping patches.

PnP-OVSS exhibits excellent performance. Comparing with MaskClip [90] and Reco [58], the two methods that require no additional training and ground truth for hyperparameter tuning, on an equal-resolution basis, we attain +29.4% mIoU on Pascal VOC, +13.2% mIoU on Pascal Context, +14.0% mIoU on COCO Object, and +11.4% mIoU on ADE-20K. Further, PnP-OVSS surpasses all baselines in Group 2. On an equal-resolution basis, we achieve +13.7% mIoU on Pascal Voc, +9.6% on Pascal Context, +9.5% on COCO Object, +11.6% on COCO Stuff, and +12.8% on ADE-20K.

When compared to Group 1, PnP-OVSS still outperforms most. On the Pascal datasets, under equal resolutions, PnP-OVSS + BLIP$_{\text{Flickr}}$ outperforms 6 out of 10 baselines on Pascal VOC, and 8 out of 9 baselines on Pascal Context. For the COCO datasets, under equal resolutions, PnP-OVSS + BLIP$_{\text{Flickr}}$ beats 5 out of 8 baselines on COCO Object, and all baselines except PACL [46] on COCO Stuff. The two Pascal datasets share many images, and the two COCO datasets use exactly the same images, but they have

Method	Finetuning VLMs	HT on Dense Labels	Short-side Resolution	Pascal VOC-20	Pascal Context-59	COCO Object-80	COCO Stuff-171	ADE 20K-150
Group 1: Methods that require weakly supervised finetuning on image-text data								
ViL-Seg† [43]	✓	✓	-	37.3	18.9	-	18.0	-
CLIPpy [51]	✓	✓	224	52.2	-	**32.0**	25.5*	13.5
SegClip [44]	✓	✓	224	52.6	24.7	26.5	-	-
GroupVit (by [51])	✓	✓	224	28.1	14.8	12.9	-	6.2
GroupVit (by [6])	✓	✓	448	50.4	18.7	27.5	15.3	9.2
GroupVit [74]	✓	✓	448	52.3	22.4	24.3	-	-
ViewCo [52]	✓	✓	448	52.4	23.0	23.5	-	-
OVSegmentor [75]	✓	✓	448	53.8	20.4	25.1	-	-
TCL [6] +PAMR [2]	✓	✓	448	55.0	30.4	31.6	22.4	17.1
PACL [46]	✓	✓	224×4	**72.3**	**50.1**	**-**	**38.8**	**31.4**
Group 2: Methods that require finetuning but not real image-text data								
MaskClip w/ ST [90]	✓	✓	336	-	**31.1**	-	**18.0**	-
MaskClip w/ ST (by [6])	✓	✓	448	38.8	23.6	20.6	16.4	9.8
ZeroSeg [8]	✓	✓	448	**40.8**	20.4	**20.2**	-	-
Group 3: Methods that require no finetuning								
MaskClip (by [51])	×	×	224	22.1	-	13.8	8.1	6.8
MaskClip [90]	×	×	336	-	25.5	-	14.6	-
Reco [58]	×	×	320	-	**27.2**	-	**27.2**	-
Reco (by [6])	×	×	448	**25.1**	19.9	**15.7**	14.8	11.2
PnP-OVSS with different VLMs								
BLIP$_{\text{Flickr}}$	×	×	224	47.8	36.4	24.8	25.8	18.2
BLIP$_{\text{Flickr}}$	×	×	320	51.7	40.4	28.0	29.6	21.3
BLIP$_{\text{Flickr}}$	×	×	336	52.5	40.7	28.2	29.6	21.9
BLIP$_{\text{Flickr}}$	×	×	448	54.5	42.2	29.7	31.5	22.6
BLIP$_{\text{Flickr}}$	×	×	768	54.1	**42.8**	31.8	32.5	**23.5**
BLIP$_{\text{COCO}}$	×	×	768	**55.7**	41.9	**33.8**	**32.6**	23.2
BridgeTower	×	×	322	36.4	32.3	24.2	27.6	18.6
BridgeTower	×	×	336	35.3	32.4	24.2	27.6	18.0
BridgeTower	×	×	770	35.2	32.4	24.5	28.0	19.0

Table 3. Zero-shot semantic segmentation performance in mIoU. Group 3 contains the most similar baselines that serve as fair comparisons to PnP-OVSS. Groups 1 and 2 benefit from additional training, extra image-text data, and hyperparameter tuning on dense labels. We use the word "by" followed by a paper citation to indicate results of the same technique reported by different papers. * CLIPpy tests on 133 categories of COCO Stuff while we test all 171 classes of COCO Stuff. ViL-Seg† is tested on subset of classes on the three datasets, as detailed in the supplementary.

Method	Dense Labels	HT on Dense Labels	Pascal VOC-20	Pascal Context-59	COCO Stuff-171
PnP-OVSS + BLIP$_{\text{Flickr}}$ (Ours)	×	×	53.6	53.8	39.8
SPNet+ST [72]	✓	✓	25.8	-	26.9
ZS3Net+ST [4]	✓	✓	21.2	20.7	10.6
CaGNet+ST [18]	✓	✓	30.3	-	13.4
STRICT [47]	✓	✓	35.6	-	30.3
LSeg [32]	✓	✓	41.0	-	-
SimBase [77]	✓	✓	72.5	-	36.3
MaskCLIP+ w/ ST [90]	✓	✓	86.1	66.7	54.7

Table 4. Comparison of zero-shot semantic segmentation performance on unseen categories with methods trained with dense annotation.

Ablated Model	Pascal Context	COCO Stuff
BLIP$_{\text{Flickr}}$	19.8	14.5
+ GradCam	21.6	17.5
+ Drop 1	25.1	19.8
+ Drop 2	26.5	20.6
+ Drop 3	27.0	20.9
+ Drop 4	27.2	20.9
+ Drop 4 + Blur	36.8	28.6
+ Drop 4 + Dense CRF	35.3	31.8
+ Drop 4 + Blur + Dense CRF	42.8	32.5

Table 5. An ablation study of PnP-OVSS + BLIP$_{\text{Flickr}}$ with resolution 768 on Pascal Context and COCO Stuff.

vastly different lists of classes to segment. We observe that the advantage of PnP-OVSS over methods in Group 1 becomes more pronounced as the number of classes on the same image increases.

When applied to BridgeTower [79], PnP-OVSS still surpasses all methods in Group 3 by 10.2% on Pascal VOC, 5.1% on Pascal Context, 8.5% on COCO Object, 0.4% on COCO Stuff, and 6.8% on ADE-20K. This showcases the plug-and-play ability of PnP-OVSS, which excels with different base networks.

We report comparisons against supervised methods in Tab 4. PnP-OVSS + BLIP$_{\text{Flickr}}$ outperforms 5 out of 7 methods on Pascal VOC, as well as every baseline except MaskCLIP+ on Pascal Context and COCO stuff. As these baselines benefit from dense supervision, the results further demonstrate the strengths of PnP-OVSS.

4.3. Ablation Study

We perform gradual ablation of the components of PnP-OVSS on BLIP$_{\text{Flickr}}$ and report the results in Tab. 5. Each component, including GradCAM, all Salience DropOut iterations, Gaussian blur, and Dense CRF, contribute positively to the final performance. In particular, the first iteration of Salience DropOut has much larger impact (+3.5/2.3) than the second iteration (+1.4/0.8), which in turn is more important than the rest. Interestingly, Gaussian blur by itself attains good performance (+9.6/7.7) whereas Dense CRF only works well when combined with blur. Dense CRF alone is worse than Gaussian blur by 1.5 mIoU on Pascal Context. This is likely caused by the fact that hard 0/1 labels resulted from thresholding are not informative unary potentials that can be leveraged by CRF effectively.

4.4. Hyperparameter Sensitivity

The choice of hyperparameters often exerts substantial influence on segmentation performance. Here we quantitatively examine how the choice of cross-attention layers and attention heads may change the segmentation mIoU on COCO Stuff. Tab. 6 shows the results obtained from the average cross-attention maps over all heads in each layer and those from different attention heads.

Layer Mean	1	2	3	4	5	6
mIoU	11.6	12.3	11.3	11.0	11.6	12.9
Layer Mean	7	8	9	10	11	12
mIoU	13.7	25.3	25.8	23.4	12.5	5.8
Head in Layer 8	1	2	3	4	5	6
mIoU	16.6	3.8	8.6	20.0	19.8	5.2
Head in Layer 8	7	8	9	10	11	12
mIoU	11.1	6.0	3.8	**29.6**	17.3	8.0

Table 6. Semantic segmentation performance using cross-attention maps averaged across all heads in a layer and separate heads in Layer 8. Results are attained with PnP-OVSS+BLIP$_{\text{Flickr}}$ on COCO Stuff and resolution 336.

We make the following observations. First, different layers and heads have drastic performance differences. The best-to-worst difference among all layers is 20, and that among heads in Layer 8 is 23.8, underscoring the importance of hyperparameter tuning. Second, the random search using the proposed reward function correctly identifies the best layer-head combination, even though Layer 8 is not the best layer based on average head performance. This indicates the effectiveness of our method.

5. Conclusions

We propose PnP-OVSS, which extracts the ability of semantic segmentation from opaque VLMs. PnP-OVSS is simple to use, requires no extra finetuning, and delivers high performance, exceeding not only all baselines that requires no finetuning, but also all baselines that do not use image-text pairs in finetuning. Its success hints at a new direction for open-vocabulary segmentation tasks leveraging large VLMs.

6. Acknowledgments

This work has been supported by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF- NRFF13-2021-0006), Singapore, and by the Vector Institute for AI, Canada CIFAR AI Chair, and NSERC Discovery Grant. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 1, 2

[2] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4253–4262, 2020. 7, 5

[3] Loris Bazzani, Alessandra Bergamo, Dragomir Anguelov, and Lorenzo Torresani. Self-taught object localization with deep networks. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 3

[4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 7, 1, 5

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6

[6] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 3, 7, 5

[7] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 2

[8] Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, and Mohamed Elhoseiny. Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 699–710, 2023. 1, 3, 6, 7, 5

[9] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4288–4298, 2022. 1

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the 16th European Conference on Computer Vision*, 2020. 2

[11] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022. 1

[12] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015. 3

[13] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multifold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2016. 3

[14] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 3

[15] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 1

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 6

[17] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 1, 3

[18] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020. 3, 7, 1, 5

[19] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023. 1

[20] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018. 1

[21] Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Jiajun Liu, Yitong Wang, Yansong Tang, Yujiu Yang, Jiashi Feng, Yao Zhao, et al. Global knowledge calibration for fast open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 797–807, 2023. 1

[22] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020. 1

[23] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom

Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 2

[25] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16886–16896, 2022. 1

[26] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. *arXiv preprint arXiv:2105.00957*, 2021. 3

[27] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 2

[28] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 695–711. Springer, 2016. 3

[29] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 3, 5

[30] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3524–3533, 2017. 3

[31] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022. 1

[32] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022. 3, 7, 1, 5

[33] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 2

[34] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 2, 3, 4

[35] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2, 3, 4

[36] Jinlong Li, Zequn Jie, Xu Wang, Xiaolin Wei, and Lin Ma. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 35:16037–16051, 2022. 1

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 2, 3, 6

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[39] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021. 2

[40] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV 2020*, 2020. 1, 2

[41] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 1, 3

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[43] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022. 2, 3, 7, 5

[44] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 1, 2, 3, 7, 5

[45] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 6

[46] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023. 1, 3, 6, 7, 5

[47] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2693–2702, 2021. 3, 7, 1, 5

[48] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19446–19455, 2023. 1

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 2, 3

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[51] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5571–5584, 2023. 2, 3, 7, 5

[52] Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*, 2023. 2, 3, 7, 5

[53] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022. 1

[54] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023. 3

[55] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023. 1

[56] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1939–1946, 2013. 3

[57] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 2, 4

[58] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35:33754–33767, 2022. 3, 6, 7, 5

[59] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021. 3

[60] Simon Blanke. Gradient-Free-Optimizers: Simple and reliable optimization with local, global, population-based and sequential techniques in numerical search spaces. `https://github.com/SimonBlanke`, since 2020. 1

[61] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, 2022. 1, 2

[62] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 3

[63] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. In *Findings of the Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*, 2022. 1, 2, 4

[64] Huy V Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8287–8296, 2019. 3

[65] Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. *Advances in Neural Information Processing Systems*, 34:16764–16778, 2021. 3

[66] Wenguan Wang, Guolei Sun, and Luc Van Gool. Looking beyond single images for weakly supervised semantic segmentation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[67] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023. 6

[68] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 3

[69] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 88:113–126, 2019. 3

[70] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 3

[71] Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. A simple baseline for knowledge-based visual question answering. In *EMNLP*, 2023. 1

[72] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 3, 7, 5

[73] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*, 2023. 2, 3

[74] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 2, 3, 7, 5

[75] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2935–2944, 2023. 1, 2, 3, 7, 5

[76] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 1

[77] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 1, 3, 7, 5

[78] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 1

[79] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. *arXiv preprint arXiv:2206.08657*, 2022. 1, 2, 3, 6, 8

[80] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *arXiv Preprint 2205.01917*, 2022. 2

[81] Nir Zabari and Yedid Hoshen. Open-vocabulary semantic segmentation using test-time distillation. In *European Conference on Computer Vision*, pages 56–72. Springer, 2022. 1

[82] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 2, 3

[83] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*, 2022. 2, 3

[84] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12765–12772, 2020. 3

[85] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 1

[86] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, 2021. 1

[87] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018. 3

[88] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 3

[89] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[90] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3, 6, 7, 1, 5

[91] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4299–4309, 2022. 1

Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Language Models

Supplementary Material

7. Additional Qualitative Results

We provide additional qualitative results of images from Pascal Context and COCO Stuff in Fig 6. We note that, in the third image of the third row, PnP-OVSS correctly recognizes four loaves of bread whereas the ground truth only annotates one.

The rightmost column of the bottom three rows, separated from the rest by a red dash line, are examples of failure cases. The example on top contains multiple small instances of the same class (`people` and `surfboard`), which are understandably hard. The middle example contains multiple instances of `people`, which causes difficulties for PnP-OVSS to cover all objects. In addition, images with a clutter of different objects and complex texture (as in the last image) often cause a drop in performance.

8. Qualitative Results in the Wild

Fig 5 show qualitative results of PnP-OVSS+BLIP$_{\text{Flickr}}$, containing objects not seen in common semantic segmentation datasets, including cartoon characters `Minions` and `Gru`, dog breeds like `Samoyed` and `Border Collie`, food items like `Hamburger`, `Fries`, `Coke`, and `Fried Chicken`, places of interest like `Eiffel Tower` and `Merry-go-round`, a new electronic vehicle `Cybertruck`, and a celebrity `Elon Mask`.

In particular, we would like to point out the difficulty involved in segmenting the Steamboat Willie image, which is in greyscale and has little texture, depriving the network the ability to use color and texture features. Despite that, PnP-OVSS is able to extract the main components of `Mickey Mouse`, the `Helm` and the `Deck`.

9. PnP-OVSS Implementation Detail

BLIP. We use the ITM branch of BLIP_large, which adopts VIT-L/16 as the image encoder, BERT as the text encoder, and insert an extra cross-attention layer for each transformer block of BERT. For each cross-attention layer, the hidden size is 768, and the number of heads is 12. We interpolate the positional embedding to allow input resolution of 768 x 768. We adopt pretrained weights from two checkpoints for image retrieval on COCO and Flickr.

BridgeTower. We use the ITM branch of BridgeTower$_large$, which adopts VIT-L/14 as the image encoder, RoBERTa$_large$ as the text encoder and an 6-layer cross attention encoder. For each cross attention layer of the cross-modal encoder, the hidden size is set to 1,024, and the number of heads is set to 16. We interpolate the positional embedding to allow input resolution of 770 x 770. The model weights are from the `bridgetowerlarge-itm-mlm-itc` checkpoint from Huggingface.

Random Search. We adopt the random search routine from the Gradient-Free-Optimizers library[1] [60] with our reward metric (§3.4). To parallelize the search process, we divide the search space into three groups and place each group on a GPU card. We perform 34 search iterations in each group. The best hyperparameter set from the three groups with the highest reward is taken as the final search result.

Class Split for Densely Supervised Models. We follow the most common setting [4, 18, 32, 47, 77, 90] which save pottedplant, sheep, sofa, train, tvmonitor as the 5 unseen classes for Pascal VOC; cow, motorbike, sofa, cat, boat, fence, bird, tvmonitor, keyboard, aeroplane as 10 unseen classes for Pascal Context; frisbee, skateboard, cardboard, carrot, scissors, suitcase, giraffe, cow, road, wallconcrete, tree, grass, river, clouds, playingfield, as 15 unseen classes for COCO Stuff.

10. Inference Speed

MaskClip, Reco, and PnP-OVSS take 0.05s, 4.41s, and 2.46s on average, respectively, for inference on a 320×320 image. We calculate the inference speeds of the three models on a single A6000 GPU with 48GB of RAM. The results are the average of 20 independent runs. Hence, PnP-OVSS achieves substantially better performance without significant increase in inference time.

11. Vil-Seg Evaluation Detail

In Tab 3, different from other methods that require weakly supervised finetuning on image-text data, Vil-Seg is evaluated on subset of datasets. Specifically, the author evaluate their method on 5 classes (potted plant, sheep, sofa, train, tv-monitor) out of the 20 object categories in PASCAL VOC; 4 classes (cow, motorbike, sofa, cat) out of the 59 object categories in PASCAL Context; and 15 classes (frisbee, skateboard, cardboard, carrot, scissors, suitcase, giraffe, cow, road, wall concrete, tree, grass, river, clouds,

[1] https://github.com/SimonBlanke/Gradient-Free-Optimizers

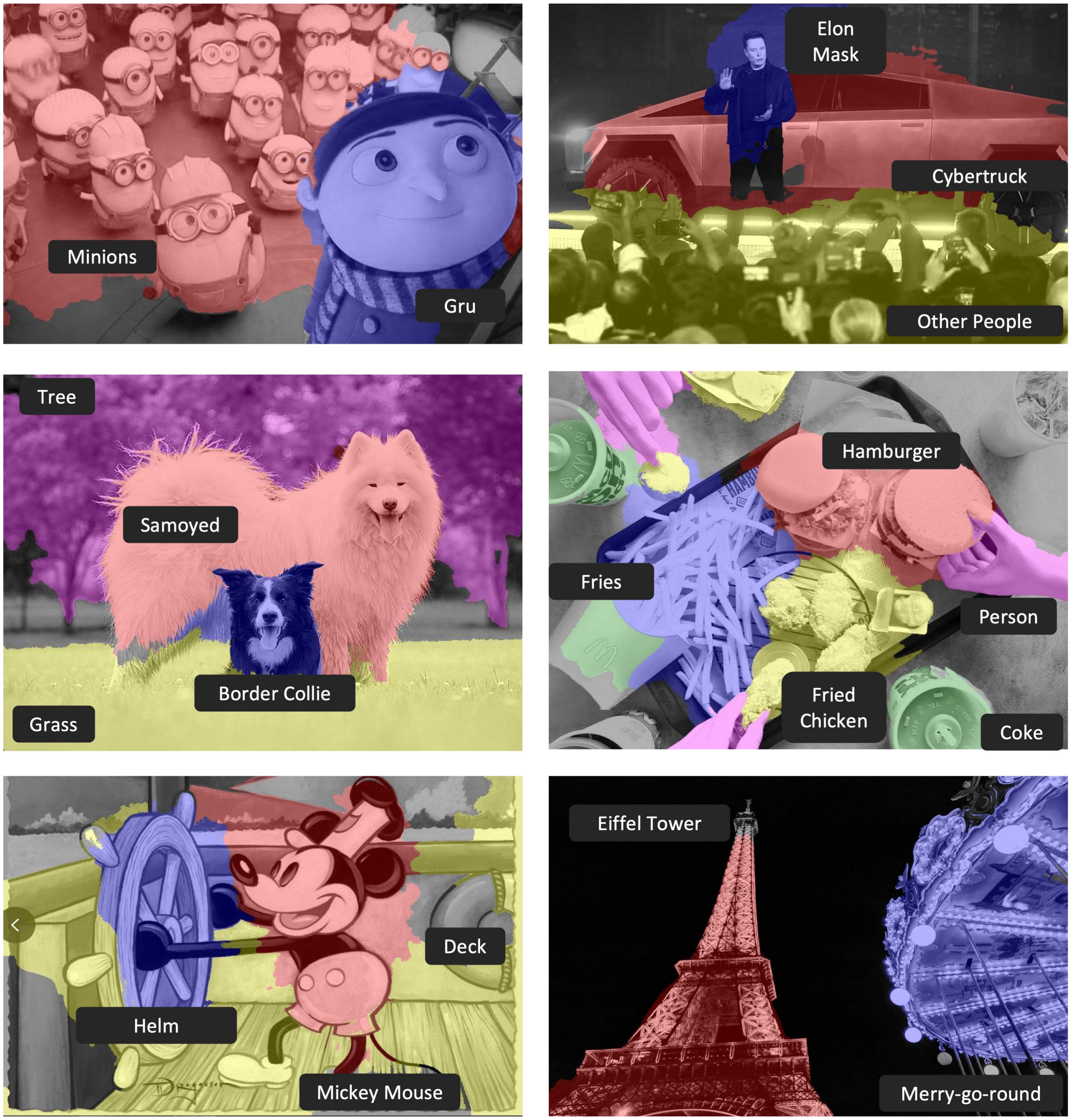


Figure 5. PnP-OVSS+BLIP$_{\text{Flickr}}$ segmentation result for in the wild images.

playing-field) out of the 171 object categories in COCO Stuff dataset.

12. Details of Zero-shot Semantic Segmentation Techniques

We summarize current methods for zero-shot semantic segmentation in Tab 7 to enable straightforward comparison between methods. Specifically, we include the supervision used, whether the method require pretraining and finetuning, the pretraining weight, finetuning data and total data size used in each method.

13. PnP-OVSS with ALBEF and mPLUG

As shown in Tab 8, we further apply PnP-OVSS on two other vision language models with cross-attention and image text matching loss, ALBEF [35] and mPLUG [34].

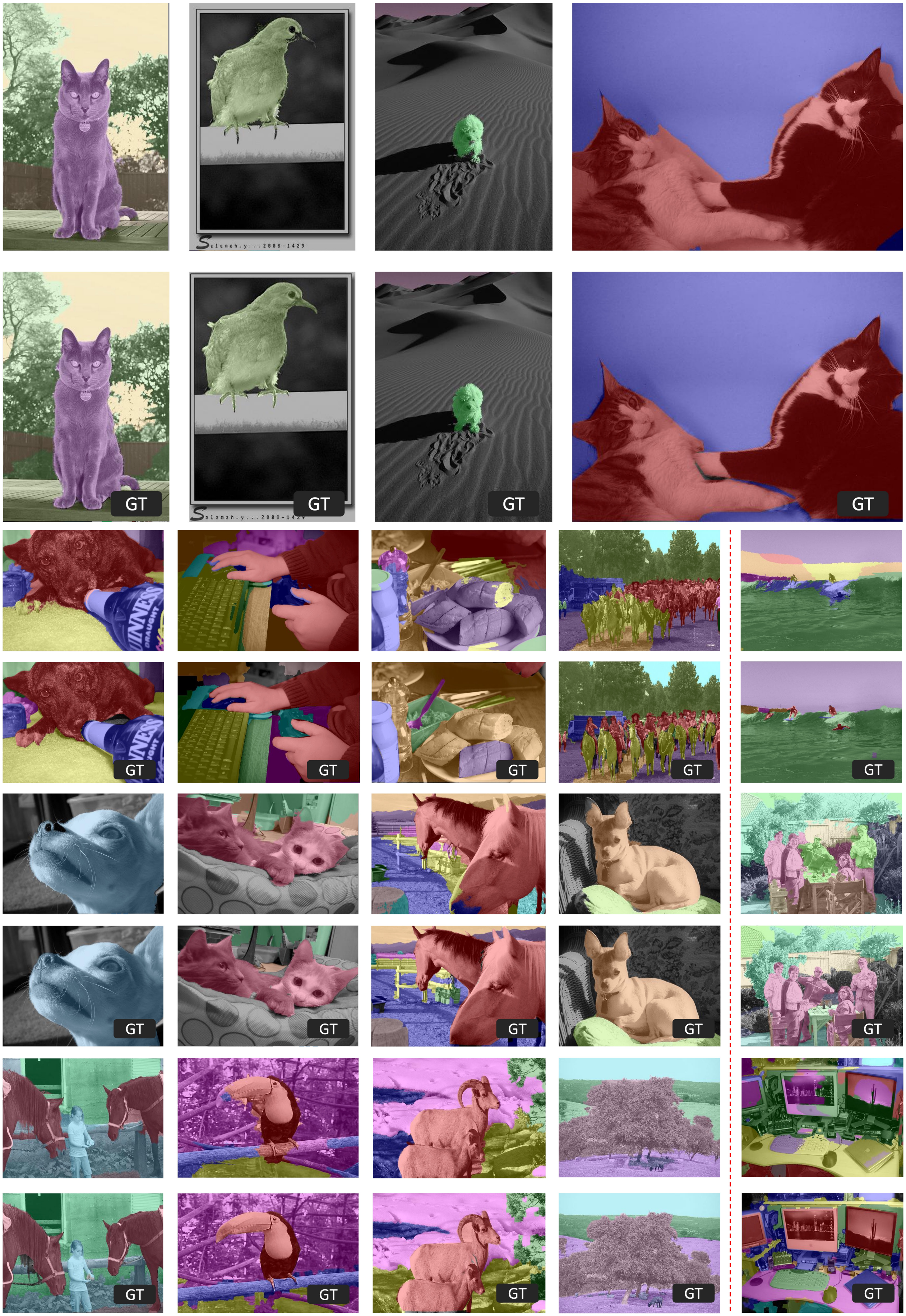


Figure 6. **Qualitative Results of PnP-OVSS + BLIP.** Images are from Pascal Context and COCO stuff. The bottom rows show the ground-truth (GT); the rest are our results. The last column of the last three rows, after the red dash line, shows failure cases.

Nevertheless, the performances are not as good as BLIP or BridgeTower. ALBEF [35] is pretrained with only 14M data with ViT-B whereas BLIP and BridgeTower are pretrained with 129M/404M data with ViT-L. We speculate that vision language models require sufficient numbers of parameters and pretraining data to acquire localization capability. mPLUG [34] is another vision language model pretrained with 14M data and ViT-L. However, mPLUG is trained for both image task and video task. With a relatively smaller amount of pretraining data than BLIP or BridgeTower, as well as image and video dual-modality objectives, mPLUG also does not perform well in image object localization.

Method	Supervision	Training	PT weight	Finetuning data	PT/T data size
Methods that require finetuning on dense annotations w/o VL Models					
SPNet [72]	Pixel+Self	PT+FT	ImageNet	Pascal Voc/COCO stuff	119K
ZS3Net [4]	Pixel+Self	PT+FT	-	Pascal Voc/ Pascal Context	5K
CaGNet [18]	Pixel+Self	PT+FT	-	Pascal Voc/Pascal Context /COCO stuff	123K
STRICT [47]	Pixel+Self	PT+FT	ImageNet	Pascal Voc/COCO stuff	119K
Methods that require finetuning on dense annotations w/ VL Models					
SimBase [77]	Pixel+Self	PT+FT	Maskformer/FCN + CLIP	COCO stuff	400.1M
LSeg [32]	Pixel+Self	PT+FT	ImageNet+CLIP	Pascal VOC/COCO/FSS	401.3M
MaskCLIP+ [90]	Pixel+Self	PT+FT	CLIP	COCO stuff	400.1M
Methods that require finetuning on image-text pairs					
OVSegmentor [75]	Text+Self	PT+FT	DINO+BERT	CC4M	4.3M
Vil-Seg [43]	Text+Self	PT+FT	CLIP	CC12M	412M
GroupVit* [44]	Text	T	-	CC3M+COCO	3.4M
GroupVit [74]	Text	T	-	CC12M+YFCC14M	26M
CLIPpy [51]	Text+Text	PT+FT	DINO+T5 Sentence	HQITP-134M	134M
SegCLIP [44]	Text+Self	PT+FT	CLIP	CC3M+COCO	403.4M
Viewco [52]	Text	PT+FT	GroupViT	CC12M+YFCC14M	26M
TCL[6]+PAMR[2]	Text	PT+FT	CLIP	CC12M+CC3M	415M
PACL [46]	Text	PT+FT	CLIP	GCC3M+GCC12M +YFCC15M	430M
Methods that require finetuning but not image-text pair					
MaskCLIP [90] w/ ST	Text+ST	PT+T	CLIP	ImageNet1K	401.2M
ZeroSeg* [8]	Text+Self	PT+T	CLIP	CC3M+COCO	403.4M
ZeroSeg [8]	Text+Self	PT+T	CLIP	ImageNet1K	401.2M
Methods that require no finetuning					
MaskCLIP [90]	Text	PT	CLIP	-	400M
Reco[58]	Text	PT	CLIP+ImageNet	-	400M
PnP-OVSS (Ours)					
+ BLIP	Text	PT	BLIP_Flickr/BLIP_COCO	-	129M
+ BridgeTower	Text	PT	BridgeTower	-	404M

Table 7. Current methods for zero-shot semantic segmentation. Pixel represents method require pixel level annotation, Self represents method leverage self supervision, and Text represents method leverage image-text pair annotation. PT stands for Pre-training, T stands for training, ST stands for Self-training, FT stands for finetuning. All methods with pixel supervision are trained on seen categories and tested on unseen categories. For data size, We calculate only the image-caption data used for pretraining and all type of data for finetuning.

Method	Training	HT on Dense Labels	Short-side Resolution	Pascal VOC-20	Pascal Context-59	COCO Object-80	COCO Stuff-171
PnP-OVSS (Ours)							
+ ALBEF	×	×	336	10.8	6.3	8.9	10.3
+ ALBEF	×	×	768	11.1	6.8	8.8	10.7
+ mPLUG	×	×	336	9.2	8.8	7.9	6.7

Table 8. Zero-shot semantic segmentation performance in mIoU.