What Holds Back Open-Vocabulary Segmentation?

Josip Šarić[1]* Ivan Martinović[2]* Matej Kristan[1] Siniša Šegvić[2] †
[1]Faculty of Computer and Information Science
University of Ljubljana, name.surname@fri.uni-lj.si
[2]Faculty of Electrical Engineering and Computing
University of Zagreb, name.surname@fer.hr

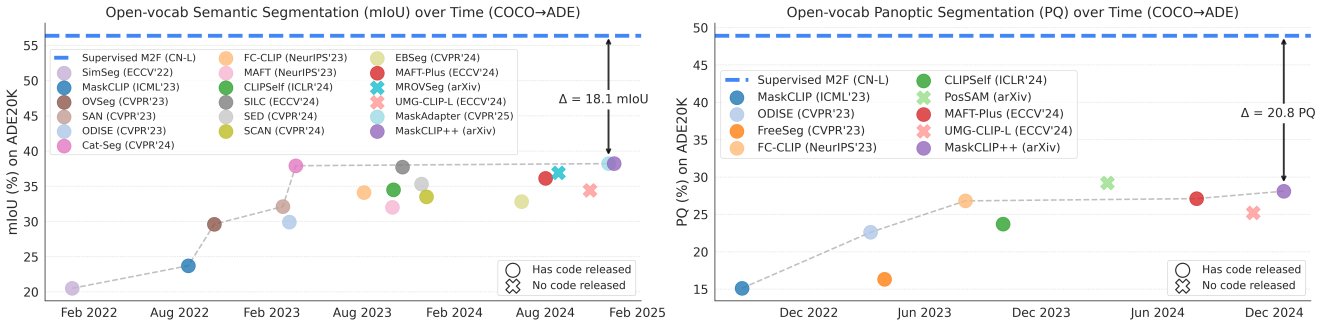

Figure 1. Performance over time of open-vocabulary semantic (left) and panoptic (right) segmentation methods (COCO→ADE20K). The blue dashed line denotes the Mask2Former supervised performance with a comparable backbone. Colored circles/crosses denote different open-vocabulary methods. We observe that the open-vocabulary methods have hit a plateau, lagging behind supervised models.

Abstract

Standard segmentation setups are unable to deliver models that can recognize concepts outside the training taxonomy. Open-vocabulary approaches promise to close this gap through language-image pretraining on billions of image-caption pairs. Unfortunately, we observe that the promise is not delivered due to several bottlenecks that have caused the performance to plateau for almost two years. This paper proposes novel oracle components that identify and decouple these bottlenecks by taking advantage of the groundtruth information. The presented validation experiments deliver important empirical findings that provide a deeper insight into the failures of open-vocabulary models and suggest prominent approaches to unlock the future research.

1. Introduction

Image segmentation is an important task in computer vision, supporting wide range of applications such as autonomous driving [9], medical imaging [45], and remote sensing [10]. The task has been well-studied under different paradigms such as semantic [38], instance [20], and panoptic segmentation [24], which unifies the former two. Despite strong progress in validation accuracy [6, 51], most of the conventional methods remain constrained to reasoning within the training taxonomy. This rigidity hinders generalization capabilities and application potential in the wild.

Recently, vision-language models (VLM) have emerged as a promising solution for recognition beyond the training taxonomy [21, 43]. Recognition based on image-text similarities enables effortless vocabulary expansion at test time by simply introducing novel class descriptions. Extending this capability to dense prediction has become a prominent focus in recent research, giving rise to the task of open-vocabulary segmentation [15, 35].

While early approaches showed promise, recent open-vocabulary segmentation models still lag significantly behind their in-domain counterparts. Figure 1 illustrates this gap for semantic (left) and panoptic segmentation (right), showing the performance of open-vocabulary models trained on COCO [36] and evaluated on ADE20K [64], alongside the closed-set Mask2Former [6] trained directly on ADE20K (blue dashed line). Notably, the best models trail the closed-set in-domain baseline by nearly 20 points. Even more concerning is the apparent stagnation in open-vocabulary performance, despite the task's relatively recent emergence and the high annotation costs of current training pipelines. In addition to relying on millions of image-caption pairs, these methods train on more than 100,000 densely annotated images from COCO. In contrast, our experiments demonstrate that in-domain models achieve com-

*Equal contribution.

parable performance with as few as 300 labeled images from ADE20K (*cf*. Fig. 4). This suggests that current open-vocabulary methods and setups share some underlying limitations that prevent them from closing the gap.

In this paper, we investigate the root causes of the exposed issue in the context of mask-transformer-based [6, 29, 51, 59] open-vocabulary segmentation methods [22, 60]. Our detailed analysis of top-performing methods identifies key bottlenecks and uncovers several insightful findings. We show that none of the core components in current open-vocabulary methods are sufficiently effective: vision-language models struggle with region-level classification, and mask proposal generators often fail to provide adequate segmentation. In fact, many valid masks are produced internally, but discarded during inference due to conflicting training and testing objectives. These findings raise concerns about the current training setup, where evaluation goals are often infeasible given the supervision provided. Finally, we outline research directions to address these issues and propel open-vocabulary segmentation forward.

2. Related Work

Image segmentation. Early deep learning approaches treated semantic and instance segmentation as separate tasks. Seminal work in semantic segmentation adapted convolutional classification networks for dense prediction [38]. Subsequent work enhanced spatial details through atrous convolutions [2], pyramidal pooling [63], and ladder-style decoders [3, 26, 45]. On the other hand, instance segmentation evolved from object detection pipelines [16, 17, 44], adding segmentation heads to produce per-instance masks [19, 20]. Later, panoptic segmentation [24] unified both tasks. Most prominent panoptic methods build on the mask transformer framework [5, 6, 50], which represents a unified architecture for all segmentation tasks.

Vision-language models. Contrastively pretrained vision–language models such as CLIP [43] and ALIGN [21] are central to open-vocabulary segmentation. Trained on large-scale data consisted of image–text pairs [4, 43, 46], they learn to embed both modalities in a shared semantic space, enabling direct cross-modal comparison. Subsequent methods enhance CLIP in three key ways: they introduce optimized classification [62] and spatial-aware localization losses [25, 48]; refine the training procedure [30, 31, 47, 61]; and curate higher-quality pre-training data [14, 54]. OpenCLIP [7] trains CLIP from scratch on the public LAION-5B dataset [46] and introduces ConvNeXt [37] backbones alongside standard ViTs [12].

Open-vocabulary segmentation. There are two main paradigms for open-vocabulary segmentation: (i) training-free [23, 27, 32, 49, 65] and (ii) training-based [11, 34, 40, 55, 58]. Training-free methods mostly rely on vision-language models (such as CLIP) to make zero-shot predictions. We focus on training-based approaches as they offer better performance and enable instance-level recognition. These approaches are further divided into (i) weakly supervised [1, 40, 55, 56] and (ii) fully supervised [8, 11, 18, 34, 53, 57, 58, 60], with the latter being more common. We focus on fully supervised methods, which train on COCO [36] and evaluate on a broad suite of test benchmarks [9, 13, 42, 64]. Fully supervised open-vocab methods aim to learn generic *objectness* from ground-truth annotations while remaining robust to the domain shift encountered at test-time. Such methods fall into two groups according to the underlying segmentation model: (i) pixel/patch-based [8] and (ii) mask-based [11, 22, 33, 34, 57, 60]. CAT-Seg [8] is the most prominent pixel-based method; it refines pixel-level cosine similarities between CLIP image and text embeddings through a cost-aggregation framework. Several subsequent works [41, 48, 52] evaluate their vision–language pretrained models within this CAT-Seg framework. Mask-based approaches classify complete masks rather than individual pixels or patches. Classification can operate in image space via mask cropping [11, 34] or in feature space through mask pooling [21, 33, 39, 57, 60]. We focus on mask-based models, which represent the most general framework as they support semantic, instance, and panoptic segmentation. Specifically, we study two recent unified models, FC-CLIP [60] and MAFT+[22], described in Section 3.

3. Preliminary

Task definition. Open-vocabulary segmentation aims to partition an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ into a set of binary masks with corresponding semantic labels:

$$\left\{(\mathbf{M}_i, c_i)\right\}_{i=1}^{N}, \qquad \mathbf{M}_i \in \{0,1\}^{H\times W}. \tag{1}$$

We denote by N the number of ground-truth segments in $\mathbf{I}$, and each segment is associated with a semantic label c_i. An open-vocabulary segmentation model is trained on a label set $\mathcal{C}_{\text{train}}$, whereas at inference time it may encounter previously unseen categories drawn from a test taxonomy $\mathcal{C}_{\text{test}}$ (*i.e*. $\mathcal{C}_{\text{train}} \neq \mathcal{C}_{\text{test}}$). We can split $\mathcal{C}_{\text{test}}$ into seen $\mathcal{C}_{\text{seen}} = \mathcal{C}_{\text{test}} \cap \mathcal{C}_{\text{train}}$ and unseen semantic categories $\mathcal{C}_{\text{unseen}} = \mathcal{C}_{\text{test}} \setminus \mathcal{C}_{\text{train}}$. We focus on standard setup with COCO as training ($\mathcal{C}_{\text{train}}$) and ADE20K as evaluation dataset ($\mathcal{C}_{\text{test}}$). A common approach in open-vocab segmentation encodes each class label as a CLIP-based text embedding and matches these embeddings against pixel- or region-level visual features.

Performance metrics. Our analysis emphasizes panoptic segmentation as the most comprehensive segmentation task. Hence, our primary metric is panoptic quality [24]:

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g)\in \text{TP}} \text{IoU}(p,g)}{|\text{TP}|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|\text{TP}|}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|}}_{\text{recognition quality (RQ)}}. \tag{2}$$

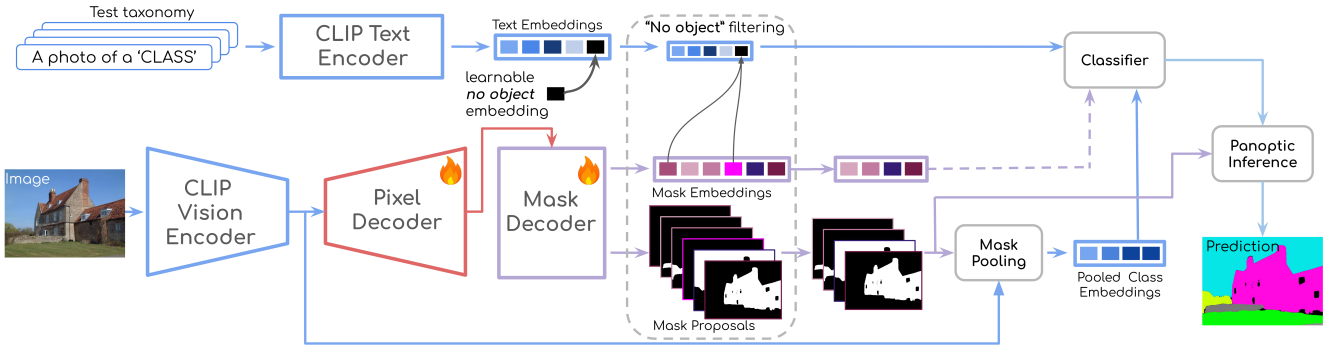

Figure 2. Standard mask-based pipeline for open-vocabulary segmentation, as proposed in FC-CLIP [60] and MAFT+ [22]. At test time, CLIP vision–text encoders supply aligned visual and textual features. A mask decoder, starting from N learnable embeddings, cross-attends to the visual features to produce class-agnostic mask proposals. Masks tagged as *no object* are discarded; the rest are labeled by ensembling a learned head (FC-CLIP) with a mask-pooled CLIP head. A panoptic inference then fuses masks and logits into the prediction.

Sets TP, FP, and FN denote the correctly matched prediction–ground truth mask-segment pairs (true positives), unmatched masks (false positives), and unmatched ground-truth segments (false negatives), respectively. A predicted mask p matches a ground-truth segment g if they share the semantic class ($c_p = c_g$) and overlap with $\mathrm{IoU}(p, g) > 0.5$. We compute PQ (*cf*. Eq. 2) independently for each class and report the mean across classes. We also report $\mathrm{PQ}_{\mathrm{seen}}$ and $\mathrm{PQ}_{\mathrm{unseen}}$ as mean PQ over the corresponding class subsets.

Open-vocab methods: FC-CLIP [60] and MAFT+ [22]. Our analysis focuses on two prominent recent open-vocabulary methods: FC-CLIP [60] and MAFT+ [22]. Both approaches follow the dominant paradigm in open-vocab segmentation and pair CLIP [43] with a mask-transformer [6]. Therefore, insights gained in this study generalize to other approaches as well. The architecture consists of three main components: i) a vision encoder that extracts features, ii) a pixel decoder that upsamples these features and iii) a mask decoder that generates mask proposals and mask embeddings (*cf*. Fig. 2). The mask decoder decouples segmentation into two distinct tasks: i) mask localization and ii) mask recognition. This design enables seamless integration of CLIP-based recognition, supporting the test-time goal of identifying unseen categories $\mathcal{C}_{\mathrm{unseen}}$ in open-vocabulary segmentation.

To this end, FC-CLIP and MAFT+ express mask-wide classification using the similarity between mask-pooled visual CLIP features and textual CLIP embeddings of the target classes. This setup can recognize previously unseen categories, provided their textual descriptions are available at test time. FC-CLIP [60] preserves vision–language alignment by freezing the CLIP encoders during training. In contrast, MAFT+ [22] fine-tunes the CLIP vision encoder but encourages the fine-tuned features to remain close to the pre-trained ones by introducing an additional loss term.

We next outline components of FC-CLIP and MAFT+ most relevant to our analysis (*cf*. Fig. 2). Let $\mathcal{V}_{\mathrm{CLIP}}$ and $\mathcal{T}_{\mathrm{CLIP}}$ denote the CLIP vision and text encoders. $\mathcal{V}_{\mathrm{CLIP}}$ extracts features $\mathbf{F} \in \mathbb{R}^{H' \times W' \times D}$ from the input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$. The mask decoder starts with N learnable embeddings, attends them to the upsampled visual features, and produces mask embeddings $\mathcal{E}_{\mathrm{mask}} \in \mathbb{R}^{N \times D}$ and pixel-to-mask scores $\mathcal{E}_{\mathrm{pixel}} \in \mathbb{R}^{N \times H \times W}$. Applying a sigmoid to $\mathcal{E}_{\mathrm{pixel}}$ yields the localization maps $\boldsymbol{\sigma} \in \mathbb{R}^{N \times H \times W}$. $\mathcal{T}_{\mathrm{CLIP}}$ generates text embeddings $\mathbf{E}_t \in \mathbb{R}^{|\mathcal{C}| \times D}$ by encoding a prompt such as `"a photo of a [class]"` for each class in $\mathcal{C}$. The textual embeddings $\mathbf{E}_t$ act as a handcrafted linear projection, that replaces the free weights in the standard M2F classifier. Note that an additional $(|C| + 1)$-th learnable *no-object* embedding $\mathbf{e}_\varnothing$ is appended to $\mathbf{E}_t$, forming the extended embedding matrix $\mathbf{E} = [\mathbf{E}_t; \mathbf{e}_\varnothing] \in \mathbb{R}^{|C|+1 \times D}$. Mask-wide classification probabilities $\mathbf{P} \in \mathbb{R}^{N \times (C+1)}$ are then obtained as $\mathcal{E}_{\mathrm{mask}} \cdot \mathbf{E}^\top$ followed by row-wise *softmax*. All masks classified as *no-object* are discarded, which leaves N' class posteriors $\mathbf{P}' \in \mathbb{R}^{N' \times C}$ and the corresponding localization maps $\boldsymbol{\sigma}' \in \mathbb{R}^{N' \times H \times W}$. Note that $\mathbf{e}_\varnothing$ is learned on the training set. This fact raises concerns about the true *openness* of these methods, and will be one of the main focuses of our analysis.

Another key component shared by FC-CLIP and MAFT+ is the mask-pooling operator, which acts on the dense CLIP features $\mathbf{F} \in \mathbb{R}^{H' \times W' \times D}$. Thresholding the localization maps $\boldsymbol{\sigma}'$ yields binarized masks $\mathbf{M} \in \{0, 1\}^{N' \times H \times W}$, with $\mathbf{M}_i = [\![\boldsymbol{\sigma}'_i \geq 0.5]\!]$. Given dense CLIP features $\mathbf{F}$ and a binary mask $\mathbf{M}_i$, mask pooling $\mathcal{MP}$ produces a mask-aggregated visual embedding $\mathbf{e}_{v_i} \in \mathbb{R}^D$:

$$\mathbf{e}_{v_i} = \mathcal{MP}(\mathbf{F}, \mathbf{M}_i) = \frac{\sum_{r,c}^{HW} \mathbf{F}[r, c, :] \cdot \mathbf{M}_i[r, c]}{\sum_{r,c}^{HW} \mathbf{M}_i[r, c]}. \quad (3)$$

For each of the N' predicted masks, CLIP-based probabilities can be computed by applying softmax (w/ the temperature τ) to the cosine similarities between visual embedding

$\mathbf{e}_{v_i} \in \mathbb{R}^D$ and textual embeddings $\mathbf{E}_t$:

$$\mathbf{p}^i_{\text{CLIP}} = \text{softmax}([\mathbf{e}^T_{v_i}\mathbf{e}_{t_1}, \mathbf{e}^T_{v_i}\mathbf{e}_{t_2}, ..., \mathbf{e}^T_{v_i}\mathbf{e}_{t_{|\mathcal{C}|}}], \tau). \quad (4)$$

As the distribution $\mathbf{P}'$ is based on training categories, open-vocab methods must incorporate the CLIP-based distribution $\mathbf{P}_{\text{CLIP}} \in \mathbb{R}^{N' \times D}$ during inference to recognize $\mathcal{C}_{\text{unseen}}$. FC-CLIP distinguishes $\mathcal{C}_{\text{seen}}$ from $\mathcal{C}_{\text{unseen}}$ and ensembles the in-vocabulary distribution $\mathbf{P}'$ with the out-vocabulary distribution $\mathbf{P}_{\text{CLIP}}$. In contrast, MAFT+ inference relies solely on $\mathbf{P}_{\text{CLIP}}$, yet obtained from the fine-tuned CLIP vision encoder. Because mask pooling can bottleneck unseen-class recognition capabilities, our analysis studies the upper bounds of this operation.

4. Empirical Analysis and Findings

Panoptic segmentation requires both accurate segmentation and correct classification to count a segment as a true positive. This requirement aligns naturally with mask transformers, where mask proposal generation (segmentation) is decoupled from the recognition (classification). This led us to ask: *how these two subtasks affect open-voc performance by themselves?* To investigate, we conduct a series of experiments in which either the segmentation or classification gets replaced with an oracle - an ideal component that performs inference using ground truth information. These oracle-based experiments quantify the upper bounds of current open-voc methods, revealing how far we could push the performance with perfect segmentation or classification.

4.1. Segmentation Oracle

We first examine CLIP's upper bound as a zero-shot classifier with oracle mask generator. Our segmentation oracle produces a set of perfect binary masks, one for each panoptic segment. We use pre-trained CLIP encoders. We extract dense image features in the shared vision-language embedding space according to the MaskCLIP [65] strategy for ViT backbones [12], while simply removing the global average pool for ConvNeXt [37]. Then, we gather per-mask representations by mask-pooling [57, 60] over the extracted CLIP features. We embedd class names into the common vector space with the CLIP text encoder. Finally, we recover panoptic segmentation by assigning each mask with the class whose text embedding yields the highest cosine similarity with the corresponding visual representation. Figure 3 illustrates the described procedure.

Table 1 presents results for base (top) and large variants (bottom) of ConvNeXt and ViT models. We evaluate three pre-trained ViT models: the original CLIP [43], OpenCLIP [7] and SigLIP2 [48]. Each of these models is tested on three resolutions: longer side 512, longer side 640, and shorter side 800 with longer side capped at 1333 [60]. All experiments interpolate the ViT positional embeddings to match the input resolution. The experiments reveal several interesting insights, which we discuss next.

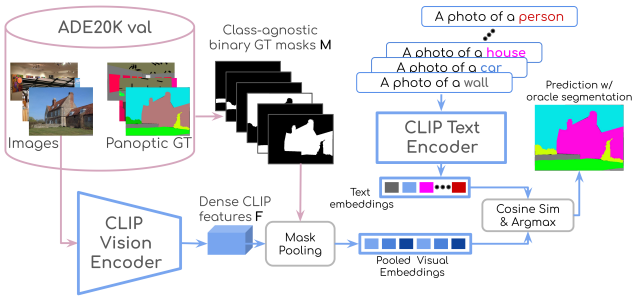

Figure 3. We estimate CLIP's out-vocab recognition ceiling using oracle masks. For each ADE20K validation image, we extract CLIP features, pool them within the ground-truth class-agnostic masks to obtain one embedding per mask, and compute cosine similarities to the CLIP text embeddings of all ADE20K classes.

First, ConvNeXt-Large achieves the highest performance among all evaluated models with 41.8 PQ. While impressive for a zero-shot setup, it falls nearly 8 points short of our in-domain Mask2Former model with the same backbone (Fig. 1). The gap is particularly striking given that the vision-language models are evaluated with oracle segmentation boundaries. These findings indicate that, despite recent advances, current VLMs still lack the dense perception required for accurate panoptic segmentation.

Second, when comparing peak performance, we find that ConvNeXt still outperforms even the most recent ViT-based model, SigLIP2. At lower input resolutions, the base variants perform comparably, and SigLIP2-Large surpasses ConvNeXt-Large. However, ConvNeXt benefits markedly from the third input configuration which employs larger and variable resolutions, while it degrades ViT models. Despite using training techniques that target dense prediction, SigLIP2 cannot match convolutional backbones. It does, however, achieve a substantial gain over earlier ViTs. This result confirms the value of its enhanced training strategy. The persistent gap suggests that ViTs still can not outperform the convolutional models at large resolutions.

Third, the performance is consistently lower on classes from ADE20K\COCO ("unseen") than on those from ADE20K∩COCO ("seen"). This gap is unexpected as our model has not received any training besides the CLIP pre-training. The result suggests that the unseen subset in open-voc segmentation experiments [60] is intrinsically harder. A plausible explanation is class-frequency bias: classes annotated in both datasets are likely more common in natural images, and VLMs may favor such frequent concepts.

Finally, we analyze the impact of capacity onto panoptic performance. ConvNeXt benefits from larger models at higher input resolutions, while showing limited gains at lower scales. In contrast, SigLIP2 demonstrates consistent

Model	Architecture	Pre-training Dataset	512×512			640×640			800×1333		
			PQ_{all}	PQ_{seen}	PQ_{unseen}	PQ_{all}	PQ_{seen}	PQ_{unseen}	PQ_{all}	PQ_{seen}	PQ_{unseen}
CLIP [43]	ViT-B-16 @ 224	WIT [43]	27.0	36.2	20.2	26.3	35.4	19.5	19.6	28.7	12.8
OpenCLIP [7]	ViT-B-16 @ 224	LAION-2B [46]	31.5	40.5	24.8	29.9	39.0	23.1	20.3	28.0	14.6
SigLIP 2 [48]	ViT-B-16 @ 256	WebLI [4]	27.6	36.9	20.7	24.2	32.5	18.0	9.9	14.7	6.3
SigLIP 2 [48]	ViT-B-16 @ 384	WebLI [4]	34.0	43.3	27.0	32.9	42.6	25.6	25.9	33.9	20.1
SigLIP 2 [48]	ViT-B-16 @ 512	WebLI [4]	33.7	43.4	26.5	34.2	44.5	26.6	30.1	39.5	23.1
OpenCLIP [7]	ConvNeXt-Base	LAION-2B [46]	32.6	41.1	26.3	35.6	45.1	28.5	39.0	50.7	30.3
CLIP [43]	ViT-L-14 @ 224	WIT [43]	12.5	19.3	7.4	13.4	19.8	8.6	13.1	20.8	7.4
OpenCLIP [7]	ViT-L-14 @ 224	LAION-2B [46]	10.8	15.9	7.0	11.2	16.7	7.0	10.7	16.3	6.5
SigLIP 2 [48]	ViT-L-16 @ 256	WebLI [4]	34.9	42.4	29.3	31.2	37.8	26.3	18.7	24.8	14.2
SigLIP 2 [48]	ViT-L-16 @ 384	WebLI [4]	38.0	45.7	32.3	36.6	44.7	30.6	27.4	34.4	22.2
SigLIP 2 [48]	ViT-L-16 @ 512	WebLI [4]	40.5	49.1	34.1	40.8	49.9	34.0	36.4	44.9	30.1
OpenCLIP [7]	ConvNeXt-Large	LAION-2B [46]	32.8	40.6	27.0	35.6	44.2	29.3	41.8	53.3	33.3

Table 1. Zero-shot panoptic quality (PQ) of dense CLIP features and our segmentation oracle on ADE20K val. We stratify classes on ADE20K∩COCO (seen) and ADE20K\COCO (unseen) according to [60].

and significant improvements across all resolutions as capacity increases. Moreover, MaskCLIP appears to extract unreliable dense features from ViT-L/14, which appears consistent with prior per-patch segmentation results [28].

Finding 1: CLIP models struggle with region-level classification and fall short of in-domain baselines even when provided with perfect segmentation.

4.2. Mask Classification Oracle

We now evaluate a classification oracle to assess the panoptic upper bound with the current mask proposals. Our oracle adjusts the class posteriors of the final panoptic map for all masks that overlap a ground truth segment with IoU > 0.5.

Table 2 presents the impact of oracle classification on open-vocabulary panoptic performance. Both FC-CLIP and MAFT+ show substantial gains of 13 PQ points overall when provided with perfect classification. Specifically, PQ_{seen} increases by 8.0 and 8.8 PQ points, while we observe a twofold improvement in PQ_{unseen}, 16.9 and 14.8 PQ points for FC-CLIP and MAFT+ respectively. These improvements confirm that recognition represents a major component of the performance bottleneck. This is particularly the case for unseen classes, which may indicate overfitting on the training dataset. However, even with perfect classification, the overall performance still falls short of typical in-domain baselines. Note that this oracle can not correct mistakes caused by the insufficient overlap between predicted and ground truth segments. This suggests that the remaining limitations lie in the quality of the mask proposals.

Model	PQ_{all}	PQ_{seen}	PQ_{unseen}
FC-CLIP	26.8	39.5	17.3
↳ + oracle classification	39.8	47.5	34.2
MAFT+	26.9	37.0	19.5
↳ + oracle classification	39.2	45.8	34.3

Table 2. Evaluating the impact of perfect mask classification on open-vocabulary panoptic segmentation on COCO→ADE20K.

Finding 2: Oracle classification improves open-voc performance but still lags behind in-domain baselines, indicating significant shortcomings in mask proposals.

4.3. Mask Selection Oracle

To better understand the root causes of the observed limitations, we dive deeper into the mask proposal generation. Our classification oracle operates only on the set of masks included in the final panoptic prediction. However, before the prediction is assembled, the mask decoder typically produces a much larger set of candidates consisting of up to $N = 250$ masks. This raises a key question: is the mask selection process that reduces this candidate set optimal?

To explore this, we conduct an experiment using an oracle mask selection. Specifically, we apply Hungarian matching between the ground truth masks and the full set of candidate masks to identify those that best explain the ground truth. The matching cost is computed using a combination of binary cross-entropy and Dice loss. After matching, we discard the unmatched candidate masks and proceed with the standard panoptic inference [6]. Note that this oracle is relatively non-intrusive, as it does not directly alter classification or segmentation, but solely influences the selection of masks from those already generated. Table 3 presents the results.

Surprisingly, oracle mask selection (*cf*. row 2) causes a significant performance drop of 4.9 PQ points and 1.5 PQ points for FC-CLIP and MAFT+. This leads us to visually inspect and compare regular predictions to those with oracle mask selection. We notice that in some cases ora-

Model	PQ_{all}	PQ_{seen}	PQ_{unseen}
FC-CLIP	26.8	39.5	17.3
↳ + oracle mask selection	21.9	33.6	13.2
↳ + dropping "no-object" logit	36.7	48.9	27.7
MAFT+	26.9	37.0	17.4
↳ + oracle mask selection	25.4	35.7	17.6
↳ + dropping "no-object" logit	33.7	42.6	27.1

Table 3. Evaluating the impact of oracle mask-selection on open-vocabulary panoptic segmentation on COCO→ADE20K.

cle selection causes disappearance of correct masks due to the filtering of masks with no-object posterior. Specifically, if the classifier assigns the highest probability to the special "no-object" class, the corresponding mask is discarded and excluded from the final panoptic segmentation. We therefore modify the inference by removing the "no-object" logit, ensuring that all oracle-selected masks are included in panoptic prediction. The third row of each section in Table 3 presents experiments with the modified selection oracle. We observe a significant performance boost of 9.9 and 6.8 PQ points for FC-CLIP and MAFT+. These findings are particularly compelling, as they suggest that open-vocabulary models could achieve significant gains through more effective mask selection alone. The improvement is especially notable on unseen classes, implying that the models internally often succeed in localizing these objects as well as assigning a correct semantic category. This indicates that many accurate masks get discarded due to poorly calibrated no-object embedding or being eclipsed by other masks. We present a more detailed analysis of this phenomena in 4.4.

Finding 3: Many valid mask proposals are discarded due to the incorrect classification as `no-object`.

Table 4 extends the analysis of the mask-selection oracle by pairing it with either oracle segmentation (row 2) or oracle classification (row 3). The segmentation oracle uses ground truth to fix the boundaries of the matched masks, while retaining the mask classifications of the corresponding open-voc segmentation model. We observe improvements of 13.6 and 18.7 PQ points over the mask-selection oracle for FC-CLIP and MAFT+, respectively. Notably, the gains are larger for the seen classes, suggesting that the models struggle more with classifying unseen categories. Additional insights emerge when comparing these results with those in Table 1, which also evaluates the segmentation oracle but in combination with the zero-shot VLMs. We observe that both FC-CLIP and MAFT+ outperform the best VLMs in presence of segmentation oracle. Interestingly, they also achieve higher performance on the unseen classes. This indicates that VLMs can benefit from ensembling with a trained mask classifier (as in FC-CLIP) or from fine-tuning of the visual backbone (as in MAFT+), even for the recognition of classes not present in the training set.

The classification oracle extends the optimal mask selection by assigning correct one-hot class probabilities based on the matching with ground truth segments. This oracle raises the performance upper bound well beyond the in-domain models. Specifically, FC-CLIP reaches 66.4 PQ, while MAFT+ follows closely with 58.1 PQ. Interestingly, performance on seen and unseen classes is comparable in this setting. This supports our earlier observation that the mask proposal generator tends to discard valid mask candidates for unseen classes, and that the classifier particularly struggles with these instances. The remaining performance gap suggests that many ground truth segments still lack appropriate corresponding masks, even with oracle mask selection. Although the matching process identifies the optimal candidate mask for each ground truth segment, it does not guarantee sufficient overlap to ensure a correct match.

+ Oracle Seg.	+ Oracle Cls.	PQ_{all}	PQ_{seen}	PQ_{unseen}
FC-CLIP w/ oracle mask selection				
–	–	36.7	48.9	27.7
✓	–	50.3	66.0	38.7
–	✓	66.4	67.1	65.9
MAFT+ w/ oracle mask selection				
–	–	33.7	42.6	27.1
✓	–	52.4	63.0	44.4
–	✓	58.1	59.8	56.8

Table 4. Evaluating the impact of oracle segmentation and classification on open-vocabulary performance on COCO→ADE20K.

Finding 4: Oracle mask proposal selection and oracle mask classification boost performance well beyond in-domain baselines, highlighting these two components as the key limitations in current open-vocabulary models.

4.4. Comparisons with In-domain Models

We complete our analysis with a more detailed performance comparison between open-vocabulary and in-domain models. Figure 4 aims to evaluate the generalization capabilities of open-vocabulary models and assess the impact of domain shift on their performance. Specifically, it presents the in-domain performance of FC-CLIP when trained on small subsets of ADE20K train and evaluated on ADE20K val, both with (yellow) and without (blue) geometric ensembling. For comparison, we also include the performance of FC-CLIP (red) and MAFT+ (green) trained on the full COCO training set. The results show that the in-domain model matches the performance of open-vocabulary models using as few as 300 labeled images. With geometric ensembling, it even surpasses them by 3 PQ points. These

findings raise questions about the practical value of open-vocabulary evaluation when comparable performance can be achieved within a target domain at such a low annotation cost. This suggests that current open-vocabulary models still struggle with domain shift. However, the domain shift between COCO and ADE20K in terms of image appearance should not be particularly pronounced, as both datasets are web-scraped from Flickr and similar sources. Hence, we argue that there must be some sort of labeling policy shift, which represents an insurmountable obstacle for current open-vocabulary models.

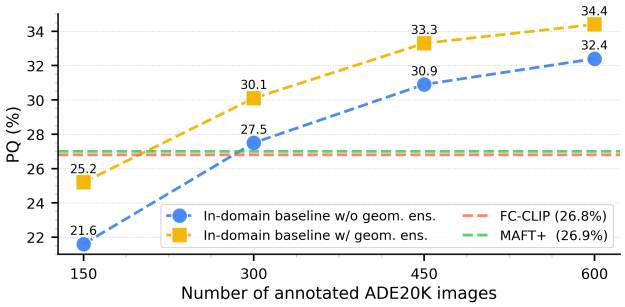

Figure 4. Performance comparison of open-vocabulary models MAFT+ (green) and FC-CLIP (red) with in-domain models trained on limited supervised data, evaluated either with (yellow) or without (blue) geometric ensembling with CLIP predictions.

To examine this more closely, we identify classes where the open-vocabulary model achieves a recall rate below 10%, and exhibits the largest drop in true positives compared to the in-domain model. Table 5 lists top five such classes, though many others are similarly affected. The table also shows the number of true positive and false negative segments. We further distinguish between false negatives caused by the absence of a predicted mask with sufficient segmentation overlap (FN_{seg}), and those where a correct mask was present but the predicted class was incorrect (FN_{cls}). We observe that the ratio of true positives to the total number of ground truth segments (FN + TP) is negligible, indicating a near-complete failure in recognizing these classes. The stratification of false negatives further reveals that most of the errors arise from inadequate segmentation rather than misclassification.

Manual inspection of COCO and ADE20K annotations reveals labeling conflicts for all classes from Table 5. These conflicts are illustrated in Figure 5. The first three columns show ADE20K images overlaid with panoptic maps from the ground truth, or predictions of FC-CLIP and MAFT+, respectively. The final column presents COCO images with overlaid panoptic maps that highlight the specific labeling conflict illustrated in each row. For clarity, some classes are omitted from the visualization. The first row illustrates the discrepancy in labeling paintings on walls. In ADE20K,

	FC-CLIP [60]			MAFT+ [22]		
class name	FN_{seg}	FN_{cls}	TP	FN_{seg}	FN_{cls}	TP
`light`	1213	27	0	1212	16	12
`painting`	708	38	40	756	20	10
`cushion`	458	16	8	472	7	3
`sign`	700	21	9	706	16	8
`pillow`	241	0	2	240	1	2

Table 5. Classes that suffer the most in transition between the in-domain and open-vocabulary setups. We show true positives (TP), and false negatives caused either by misssegmentation (FN_{seg}) or missclassification (FN_{cls}).

paintings are labeled as a distinct thing class, `painting, picture`, whereas in COCO, they are most often left unlabeled or occasionally treated as part of the surrounding wall segments. As a result, both FC-CLIP and MAFT+ fail to recognize paintings in ADE20K. This failure stems from the limitations of the mask proposal generator, which is trained with a supervision that directly contradicts the evaluation objectives. Specifically, since paintings are not labeled in COCO, the model is never encouraged to generate masks that cover painting regions. If such a mask is proposed by chance, the model is trained to classify it as `no-object`. Consequently, most of these masks are discarded before the VLM even has an opportunity to classify them as paintings. The second row considers the class `pillow`, which is present in the taxonomies of both datasets. However, COCO excludes sleeping pillows on beds and labels them as part of the bed. On the other hand, ADE20K uses `pillow` specifically for sleeping pillows and labels other types as `cushion`. As a result, open-vocabulary models trained on COCO fail to recognize bed pillows in ADE20K, as shown in the first two rows. This example illustrates the ambiguity of class names and the necessity for more accurate class descriptions in open-vocabulary segmentation. This also highlights another limitation of current open-voc models: they struggle to recognize classes that are subparts of training categories, as mask proposal generators are trained to produce a single mask for the whole object. The third row presents an example for the class `signboard, sign`, which in ADE20K includes traffic signs. In contrast, COCO provides a dedicated label only for stop signs, while the other traffic signs are often left unlabeled or annotated as part of surrounding objects, such as buildings. This inconsistency leads to similar recognition issues as before.

Finding 5: Annotation conflicts between COCO and ADE20K reveal a misalignment between training supervision and evaluation objectives, causing open-vocabulary models to discard valid mask proposals.

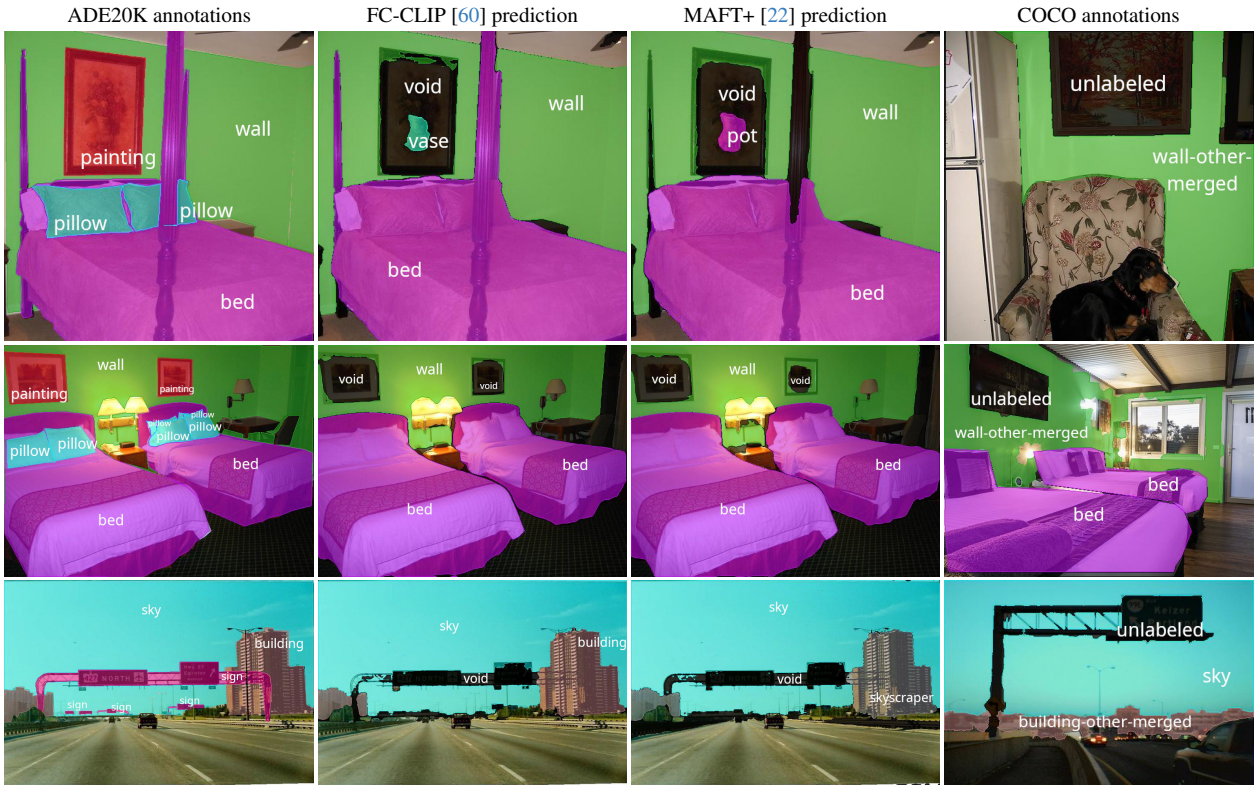

Figure 5. Illustration of labeling policy conflicts between COCO and ADE20K that hinder open-vocabulary model performance. The rows illustrate labeling conflicts for classes `painting`, `pillow`, and traffic `sign`, respectively.

5. Conclusion

We have presented a comprehensive analysis of mask-transformer methods for open-vocabulary segmentation. Our oracle experiments identify the key bottlenecks in both panoptic subtasks: mask classification and segmentation. First, the current vision–language models struggle with region-level classification: even with perfect masks, they lag behind in-domain performance. This gap underscores the need for better VLM pre-training pipelines to enhance their dense representations. Second, our analysis shows that even with oracle classification, prominent open-vocabulary approaches still lag behind in-domain models, revealing shortcomings in mask proposal generation. Third, we demonstrate that mask proposal generators internally produce valid proposals, yet discard them at inference due to biases inherited from the training data. These biases prevent them from reconciling labeling-policy discrepancies between training and evaluation datasets.

The identified bottlenecks suggest several promising research directions. First, future work should eliminate the taxonomy conflicts by unifying or precisely mapping label sets in benchmark design to ensure fair and reliable evaluation of open-vocabulary performance. Second, current proposal generators lack vocabulary awareness: they generate the same mask candidates regardless of the test-time taxonomy. Future work should develop vocabulary-aware proposal generators that dynamically adapt mask boundaries to the evaluation vocabulary. Finally, we argue that richer annotation guidance is essential and can be achieved in two ways: i) in a few-shot setting, by providing exemplar segmentations that illustrate the desired outputs, and ii) by supplying detailed textual guidelines that define annotation rules for each semantic category. The latter approach appears more feasible, as articulating annotation guidelines in natural language tends to be easier than collecting examples that cover every edge case. However, language-based approach could be limited because the current CLIP text encoders operate largely as a bag-of-words. Replacing these encoders with large language models could capture finer class distinctions, but only if the vision encoders can match such subtle differences. We believe these directions offer a roadmap toward more *open* open-vocab segmentation.

Acknowledgments

This work has been supported by Croatian Recovery and Resilience Fund - NextGenerationEU (grant C1.4 R5-

I2.01.0001), Advanced computing service provided by the University of Zagreb University Computing Centre - SRCE, Slovenian research agency research program P2-0214, and European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Postdoctoral Fellowship Programme, SMASH co-funded under the grant agreement No. 101081355. The SMASH project is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund. In memory of dear mentor, colleague and friend, Siniša Šegvić.

References

[1] 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. 2

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[4] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2, 5

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 2

[6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 1, 2, 3, 5

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 2, 4, 5

[8] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2

[10] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018. 1

[11] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11583–11592, 2022. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 4

[13] 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. 2

[14] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[15] 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

[16] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[18] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1086–1096, 2023. 2

[19] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 297–312. Springer, 2014. 2

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[21] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2

[22] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 399–416. Springer, 2024. 2, 3, 7, 8

[23] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 299–317. Springer, 2024. 2

[24] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 1, 2

[25] Kevis kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, Dan Gnanapragasam, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. TIPS: Text-image pretraining with spatial awareness. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[26] Ivan Kreso, Sinisa Segvic, and Josip Krapac. Ladder-style densenets for semantic segmentation of large natural images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 238–245, 2017. 2

[27] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024. 2

[28] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024. 5

[29] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023. 2

[30] Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training. *Advances in Neural Information Processing Systems*, 36:49068–49087, 2023. 2

[31] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23390–23400, 2023. 2

[32] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv e-prints*, pages arXiv–2304, 2023. 2

[33] Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14998–15008, 2025. 2

[34] 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. 2

[35] 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

[36] 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 *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 1, 2

[37] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2, 4

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[39] Ivan Martinović, Josip Šarić, Marin Oršić, Matej Kristan, and Siniša Šegvić. Dearli: Decoupled enhancement of recognition and localization for semi-supervised panoptic segmentation. *arXiv preprint arXiv:2507.10118*, 2025. 2

[40] 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. 2

[41] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2

[42] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 2

[43] 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. 1, 2, 3, 4, 5

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer, 2015. 1, 2

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 5

[47] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[48] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 4, 5

[49] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024. 2

[50] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021. 2

[51] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419, 2023. 1, 2

[52] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 2

[53] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3426–3436, 2024. 2

[54] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 2

[55] 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

[56] 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 (CVPR)*, pages 2935–2944, 2023. 2

[57] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2, 4

[58] 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. 2

[59] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022. 2

[60] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 2, 3, 4, 5, 7, 8

[61] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022. 2

[62] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[63] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[64] 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 (CVPR)*, pages 5122–5130, 2017. 1, 2

[65] 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. 2, 4