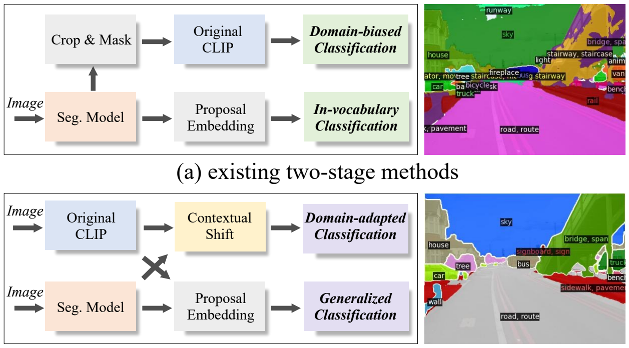
Open-Vocabulary Segmentation with Semantic-Assisted Calibration

Yong Liu[1]*, Sule Bai[1]*, Guanbin Li[2] , Yitong Wang[3] , Yansong Tang[1]†
[1]Shenzhen Key Laboratory of Ubiquitous Data Enabling, Shenzhen International Graduate School, Tsinghua University, China [2]Sun Yat-sen University [3]ByteDance Inc.
{liuyong23, bsl23}@mails.tsinghua.edu.cn, tang.yansong@sz.tsinghua.edu.cn

Abstract

This paper studies open-vocabulary segmentation (OVS) through calibrating in-vocabulary and domain-biased embedding space with generalized contextual prior of CLIP. As the core of open-vocabulary understanding, alignment of visual content with the semantics of unbounded text has become the bottleneck of this field. To address this challenge, recent works propose to utilize CLIP as an additional classifier and aggregate model predictions with CLIP classification results. Despite their remarkable progress, performance of OVS methods in relevant scenarios is still unsatisfactory compared with supervised counterparts. We attribute this to the in-vocabulary embedding and domain-biased CLIP prediction. To this end, we present a Semantic-assisted CAlibration Network (SCAN). In SCAN, we incorporate generalized semantic prior of CLIP into proposal embedding to avoid collapsing on known categories. Besides, a contextual shift strategy is applied to mitigate the lack of global context and unnatural background noise. With above designs, SCAN achieves state-of-the-art performance on all popular open-vocabulary segmentation benchmarks. Furthermore, we also focus on the problem of existing evaluation system that ignores semantic duplication across categories, and propose a new metric called Semantic-Guided IoU (SG-IoU). Code is available here.



(a) existing two-stage methods

(b) our semantic-assisted calibration network

Figure 1. Illustration of existing two-stage methods and our SCAN. Limited by domain-biased CLIP classification and in-vocabulary model classification, existing methods struggle to align visual content with unbounded text. By incorporating generalized semantic guidance of CLIP to proposal embedding and perform contextual shift, our SCAN achieves excellent OVS performance.

1. Introduction

Semantic segmentation is one of the most fundamental tasks in computer vision, which targets at assigning semantic category to pixels in an image. Despite achieving excellent performance in recent years [2, 11, 13, 17, 29, 30, 37, 40, 46], traditional semantic segmentation approaches rely on predefined sets of training categories. Consequently, these methods falter when encountering categories absent during the training phase, significantly impeding their real-world applicability. Such challenge has inspired the exploration of Open-Vocabulary Segmentation (OVS) setting [8, 12, 14, 16, 22, 31, 45, 53]. Different from traditional closed-set segmentation, OVS methods can segment arbitrary categories given only text inputs as guidance, which has many potential applications such as autonomous navigation for UAV [4, 42] and UGV [20], mobile crowdsensing [5, 26], large scale robot swarm collaborations [39].

It is extremely challenging to accurately identify unseen categories without the intervention of external knowledge. Therefore, an intuitive idea is to introduce large-scale vision-language model [19, 38] trained with numerous sources to extend the semantic space of segmentation models. Motivated by this, some studies [8, 12, 48, 49] adopt a two-stage pipeline. This approach first generates class-agnostic mask proposals with segmentation models, following which the pre-trained CLIP [38] serves as an additional classifier to execute mask-level visual-linguistic alignment. The objective is to recognize open-vocabulary concepts by combining the prior knowledge of both CLIP and segmentation models. Despite advancements under this paradigm, its capacity to align visual content with unbounded text

[1] Equal contribution
[2] Corresponding author

still falls below the anticipated outcomes considerably. As shown in Figure 1 (a), we analyze this contrast mainly stems from two aspects: 1) the proposal embedding of segmentation model is turned to fit training semantic space, making segmentation model classification collapse into in-vocabulary prediction and insensitive to novel concepts. 2) the visual inputs for pre-trained CLIP have significant domain bias. Specifically, to highlight the target area and mitigate the influence from undesired regions, the input to CLIP is sub-images after cropping and masking, which deviates significantly from the natural image distribution, *i.e.*, the visual domain of pre-trained CLIP. Such bias leads to the loss of contextual guidance as well as incorrect background prior, and thus impairs the performance.

Therefore, a natural question arises: how to introduce unrestricted knowledge space while mitigating domain bias caused by unnatural background and providing global context? It occurred to us that CLIP has well-aligned visual-linguistic space and strong capability of detecting latent semantics from natural images. The `[CLS]` token embedding extracted by CLIP condenses the context of the whole image and implicitly expresses the associated semantic distribution. With this semantic assistance, feature space of proposal embedding and the biased visual domain of CLIP can be calibrated towards more generalized recognition.

Inspired by this, we present a Semantic-assisted CAlibration Network (SCAN). On the one hand, SCAN employs a semantic integration module designed to incorporate the global semantic perception of original CLIP into proposal embedding. It extends the semantic space and alleviates the potential degradation towards in-vocabulary classification. On the other hand, we propose a contextual shift strategy to advance the open-vocabulary recognition ability of CLIP for domain-biased images. By replacing background tokens with appropriate contextual representations, *i.e.*, `[CLS]` embedding of whole image, this strategy mitigates domain bias at the feature level through semantic complementation. With above designs calibrating both in-vocabulary and out-vocabulary semantic space, our SCAN achieves the best performance on all popular open-vocabulary semantic segmentation benchmarks. Extensive experiments and analysis also demonstrate the rationality of our motivation and proposed modules.

Apart from the above contribution, we also focus on the problem of current evaluation system that neglects semantic relationships among different categories. For example, "table" and "coffee table" exist in ADE20K-150 [54] dataset as different class simultaneously and the model needs to accurately distinguish between them. If a model assigns "table" tag to a region whose ground truth label is "coffee table", it will be considered incorrect. We believe that under open-vocabulary scenarios, correct recognition of general semantics is sufficient, and there is no need to make this level of detailed hierarchical distinction. To this end, we present a new metric called Semantic-Guided IoU (SG-IoU), which takes semantic relationships between different categories into account during IoU calculation.

Our contributions can be summarized as follows:

- We present a Semantic-assisted Calibration Network (SCAN) to boost the alignment between visual content with unbounded linguistic concepts and thus improve open-vocabulary segmentation performance.
- We propose semantic integration module to alleviate in-vocabulary degradation of proposal embedding assisted by original CLIP. Besides, contextual shift strategy is applied to achieve domain-adapted alignment, mitigating the lack of global context and invalid background noise.
- We propose a new evaluation metric called Semantic-Guided IoU (SG-IoU). It takes the semantic relationships of different categories into account, which is more compatible with the open-vocabulary setting.
- Our SCAN achieves state-of-the-art on all popular benchmarks with both vanilla mIoU and our proposed SG-IoU as metric. Extensive experiments are conducted to prove the effectiveness and rationality of the proposed modules.

2. Related Work

Open-Vocabulary Segmentation. The open-vocabulary segmentation task aims to segment an image and identify regions with arbitrary text queries [1, 12, 34, 45]. Pioneering work [45] replaces the output convolution layer by computing the similarity between visual features and linguistic embeddings, which has become common practice. More recently, a two-stage pipeline [8, 12, 27, 48, 49] is proposed: the model first generates class-agnostic mask proposals, then a pretrained CLIP [38] is utilized to perform sub-image classification by cropping and masking corresponding regions. Afterward, the prediction of CLIP is ensembled with the classification results of segmentation model. With the combination of both in-vocabulary and out-vocabulary classification, these methods obtains excellent improvement. Subsequently, SAN [50] designs a side-adapter network to leverage CLIP features for decoupling segmentation and classification. OVSeg [27] observes that masked background regions affect the recognization ability of CLIP due to the distribution difference. Thus, it proposes to finetune the pretrained CLIP with such images and collect a domain-biased training dataset. Aiming to improve model efficiency, GKC [14] presents text-guided knowledge distillation strategy to transfer CLIP knowledge to specific classification layer. Although above methods have made remarkable progress, they are still susceptible to bounded training semantic space due to the crucial learnable part exists in framework. To address the overfitting and domain-biased problems, we propose SCAN that calibrates both in-vocabulary and out-vocabulary space with the assistance of

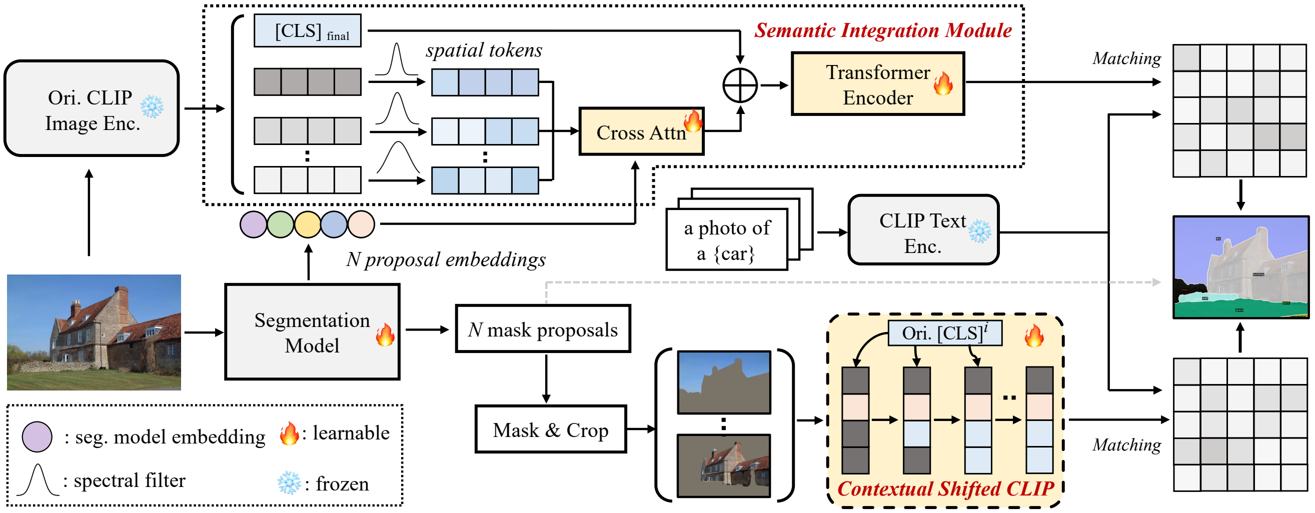


Figure 2. Pipeline of SCAN. Firstly, a segmentation model is used to generate class-agnostic masks and corresponding proposal embeddings for cross-modal alignment. To avoid collapse into known categories, the proposal embeddings are calibrated by integrating global semantic prior of CLIP in Semantic Integration Module. Besides, the cropped and masked images are input to Contextual Shifted CLIP for domain-adapted classification. Finally, the matching scores of both model embeddings and CLIP are combined to assign category labels.

global semantic prior of CLIP.

Vision-Language Pre-training. Vision-language pre-training aims to learn a joint visual-linguistic representation space. Limited to small-scale datasets, early approaches [6, 23, 25, 33, 36] struggled to achieve good performance and required fine-tuning on downstream tasks. With the availability of large-scale web data, recent works [18, 38] have showed the advantages of leveraging such data to train a more robust multi-modal representation space. Among them, CLIP [38], the most popular vision-language method, leverages the idea of contrastive learning to connect images with their corresponding captions and has achieved impressive cross-modal alignment performance. Inspired by previous works [8, 22, 24, 49], we also take advantage of the well-aligned and generalised space of CLIP to enhance open-vocabulary segmenation.

3. Method

Figure 2 shows the pipeline of SCAN. The framework follows two-stage paradigm, *i.e.*, we first take a segmentation model [7] to generate a group of class-agnostic mask proposals $M_N \in \mathbb{R}^{N\times H\times W}$ and corresponding proposal embeddings $F_N \in \mathbb{R}^{N\times C}$, where N and C indicate the number of learnable queries and embedding dimension. H, W denote the spatial size of input image. The proposal embeddings are leveraged to align with linguistic features for model-classification. In our SCAN, a Semantic Integration Module (SIM) is proposed to transfer global semantic prior originated from CLIP into proposal embeddings F_N, which calibrates the model feature space to accommodate both in-vocabulary and out-vocabulary semantics. On the other hand, mask proposals M_N are used to generate sub-images by cropping and masking related regions from input natural image. The processed sub-images are sent to CLIP [38] for classification at the mask level. We propose a Contextual Shift strategy (CS) to alleviate the domain bias and noise caused by masked background pixels and improve the classification performance of CLIP for such sub-images. Finally, the classification results from both CLIP and proposal embeddings are combined for output.

3.1. Semantic Integration Module

The learnable proposal embedding used for model classification suffers from overfitting to training semantics and insensitive to novel categories. To relieve this problem, we propose the Semantic Integration Module (SIM). The core idea of SIM is to calibrate the semantic response of mask proposal embeddings by incorporating the prior knowledge of CLIP [38]. In SIM, we use a frozen CLIP to extract implicit semantics of the input image $I \in \mathbb{R}^{H\times W\times 3}$ and obtain the progressive features $\{F^i_{HW}, F^i_{CLS}\}$, where $F^i_{HW} \in \mathbb{R}^{\frac{H}{14}\times\frac{W}{14}\times C}$ and $F^i_{CLS} \in \mathbb{R}^{1\times C}$ denote the output of *i-th* layer in CLIP image encoder. To fully utilize the coarse-grained and fine-grained perception of CLIP, we introduce both the spatial tokens F_{HW} and the general `[CLS]` token F_{CLS} into proposal embeddings.

Considering that the purpose of feature integration is to benefit high-level semantic matching, directly interacting with spatial token embedding F_{HW} may bring harmful texture noise due to local details involved in F_{HW}. Some theoretical researches [43, 51, 52] about neural network from spectral domain propose that low-frequency components correspond to high-level semantic information while ignoring details. Inspired by that, we design a simple low-frequency enhancement structure to suppress potential noise. Take F^i_{HW} as an example, the process of performing low-frequency enhancement can be represented by:

$$\begin{aligned} g^i &= Gaussian(h, w, \sigma), \\ F^i_s &= FFT(F^i_{HW}) * g, \\ F^i_s &= IFFT(ReLU(Conv(F^i_s))) + F^i_{HW}, \end{aligned} \tag{1}$$

where FFT and $IFFT$ are Fourier transform and Fourier inverse transform. g^i denotes the filtering coefficient map with the same spatial size of the feature F^i_{HW}. The center of the coefficient map has the value of 0 and increases around in the form of Gaussian (without spectrum centralization, the center of the spectrum after FFT is high frequency, and the surrounding is low frequency). σ is cutoff frequency and $*$ means element-wise product.

After performing low-frequency enhancement on selected CLIP layers, we concatenate the enhanced features to $F_s \in \mathbb{R}^{m \times \frac{H}{14} \times \frac{W}{14} \times C}$, where m denotes the number of selected CLIP layers. Then, the content prior of F_s is injected to proposal embeddings F_N by multi-head cross-attention with F_N as the Query and F_s as the Key and Value:

$$F'_N = MHA(F_N, F_s, F_s), \tag{2}$$

where MHA indicates vanilla multi-head attention and $F'_N \in \mathbb{R}^{N \times C}$, N is the number of learnable query.

The role of proposal embedding F'_N is to align with unbounded linguistic features, while the spatial tokens in the middle layer of CLIP have not actually been transformed into the vision-language unified space. Thus, we further leverage the aligned CLIP visual embedding $F^{final}_{CLS} \in \mathbb{R}^{1 \times C}$ to bridge the gap between visual and linguistic space. We add the F^{final}_{CLS} to F'_N with a learnable factor γ initialized as 0.1 under the help of broadcast mechanism. Then the features are fully interacted in transformer encoder layer and generate the final aligned proposal embeddings F^{final}_N. This process can be formulated as:

$$F^{final}_N = Trans.Enc.(F'_N + \gamma * F^{final}_{CLS}). \tag{3}$$

3.2. Contextual Shift

By taking pre-trained CLIP [38] as an extra classifier, previous two-stage approaches expect to exploit the powerful generalization capability of CLIP to tackle novel classes. But the reality is not as perfect as it seems. As Figure 3 shows, the image domain has been greatly shifted from natural distribution due to the masked patches. Such domain bias, coupled with the lack of global context, can dramatically deteriorate the recognition ability of CLIP. Besides, the shape of masked pixels also interferes with CLIP predictions. Masking background forces corresponding regions to the same value, which imposes strong erroneous prior. For example, a plane in the sky is masked. The shape of the plane still causes related response while it is difficult to recognize the concept of "sky" in the foreground due to its irregularity and unnatural background.

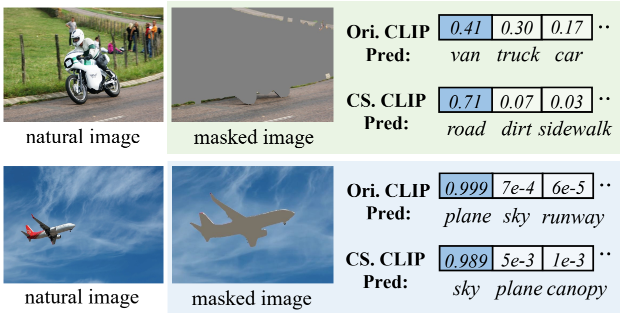


Figure 3. Illustration of image domain bias and corresponding detriment to vision-language alignment. The right side shows the classification confidence for masked images. "Ori.CLIP" and "CS.CLIP" demonstrate the original CLIP and our contextual shifted CLIP, respectively.

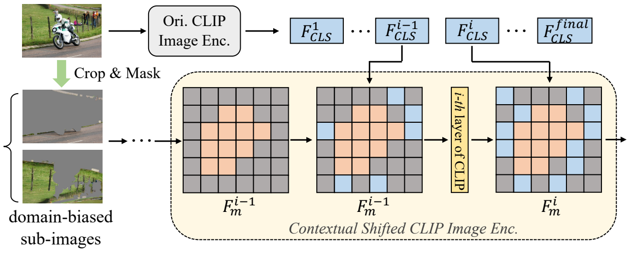


Figure 4. Process of applying contextual shift strategy.

To address this issue, we propose the Contextual Shift (CS) strategy. The key idea of CS is to replace the background token embeddings with global `[CLS]` token generated by original CLIP from whole image during the forward process. Considering the different sizes and shapes of various segmentation masks, we randomly replace a certain percentage α of the background areas in selected layers of CLIP. Take the k-th segmentation masks as example, the vanilla forward process of CLIP for region classification is:

$$F^i_m = \begin{cases} \mathcal{V}^i(\delta(I, M_k)), & \text{if } i = 0 \\ \mathcal{V}^i(F^{i-1}_m), & \text{if } i \geq 1 \end{cases} \tag{4}$$

where $\mathcal{V}^i$ denotes the i-th layer of CLIP visual encoder. F^i_m is the output features of i-th layer. δ indicates the crop and mask operation for generating sub-images. M_k and I are segmentation mask and original input image, respectively. With the CS strategy that introducing context prior within global `[CLS]` token embedding generated by original CLIP from natural image, the updated forward process can be for-

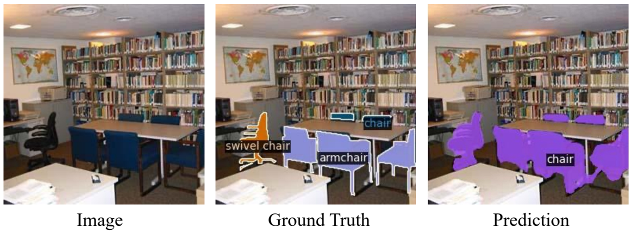


Image Ground Truth Prediction

Figure 5. Explanation of potential problems exist in the current evaluation system. There exists severe semantic duplication, *i.e.*, synonyms and parent categories, in benchmarks, while current metric does not take the semantic relationships between different categories into account.

mulated as:

$$F_m^i = \begin{cases} \mathcal{V}^i(\delta(I, M_k)), & \text{if } i = 0 \\ \mathcal{V}^i(F_m^{i-1}|(M_k, F_{CLS}^{i-1}, \alpha)), & \text{if } i \in idx \\ \mathcal{V}^i(F_m^{i-1}), & \text{if } others \end{cases} \tag{5}$$

where F_{CLS}^{i-1} is `[CLS]` embedding generated by (i-1)-th CLIP layer from natural input image. idx indicates the selected replacing layers of CLIP. $(F_m^{i-1}|(M_k, F_{CLS}^{i-1}, \alpha))$ means replace $\alpha\%$ of the mask background area with the `[CLS]` embedding F_{CLS} extracted from original image. The process is illustrated in Figure 4.

On the one hand, the global `[CLS]` embedding obtained from natural image can provide contextual information to relieve the domain bias and aid semantic prediction. On the other hand, such random replacing operation disrupts the shape of the background area, reducing the effect of error distribution of consistent background pixels. As shown in Figure 3, CS strategy can greatly improve the cross-modal alignment of domain-biased images with the aforementioned advantages. Besides, to better adapt CLIP to the shifted domain, we also follow OVSeg [27] to finetune the contextual shifted CLIP on the masked images dataset [27]. The dataset is collected from COCO Caption [3].

3.3. Semantic-Guided Evaluation Metric

Existing OVS works tend to directly take supervised semantic segmentation benchmarks with mIoU metric for evaluation. However, we observe that such evaluation is not completely applicable to open-vocabulary settings. Specifically, there exists severe semantic duplication, *i.e.*, synonyms or hypernyms, in these supervised benchmarks. For example, "chair", "armchair", and "swivel chair" exist in the ADE20K-150 dataset [54] as different class simultaneously. As Figure 5 shows, if a model assigns "chair" tag to a region but the corresponding ground truth label is "armchair", it will be considered incorrect in the existing evaluation system. Such category setting and evaluation is appropriate for closed-set segmentation tasks because their models are trained to distinguish between these fine-grained concepts. But for open-vocabulary segmentation setting, we argue that the responsibility of the model is to discern the correct semantic, *e.g.*, it should also be correct if models recognize the regions belong to "armchair" as "chair". In addition, since the required categories are manually given under real scenarios, users will not be inclined to give semantic duplicated categories.

Inspired by this observation, we propose to reorganize the calculation process of mIoU under existing popular benchmarks and present a new metric called Semantic-Guided IoU (SG-IoU) for open-vocabulary setting. The core idea of SG-IoU is to take semantic relationships between different categories into account when calculating whether a prediction is consistent with the ground truth. Specifically, we manually determine the hierarchical relations among various categories and obtain a series of category semantic association matrix. When calculating the intersection between prediction and ground truth, regions predicted to be corresponding parent or synonymous classes are also taken into account. In addition, we employ a balance factor to avoid erroneous metric boosts due to the potential overfavouritism of the model to the parent categories. This factor is related to the accuracy of the parent classes. Take q-th class as an example, the calculation process can be formulated as:

$$SG\text{-}IoU(q) = \frac{P_qG_q + P_QG_q * \beta}{P_q + G_q - P_qG_q}, \beta = \frac{P_QG_q + P_QG_Q}{P_Q} \tag{6}$$

where P_QG_q means the predicted class is Q and the ground truth category is q. Q is the synonyms and parent categories of q. β is the balance factor. Due to limited space, please see more descriptions and demonstration of the category semantic association in supplementary materials.

4. Experiments

4.1. Datasets and Evaluation Metrics

Training Dataset Following previous works [8, 21, 27, 49, 50], we train the segmentation model of our SCAN on COCO-Stuff [28] with 171 categories. CLIP [38] for mask-level classification is finetuned on masked images dataset proposed by OVSeg [27]. The dataset is collected from COCO Captions [3].

Evaluation Dataset To evaluate the effectiveness of our method, we conduct extensive experiments on the popular image benchmarks, ADE20K150 [54], ADE20K847 [54], Pascal VOC [10], Pascal Context-59 [35], and Pascal Context-459 [35].

ADE20K is a large-scale scene understanding benchmark, containing 20k training images, 2k validation images, and 3k testing images. There are two splits of

Table 1. Performance comparison with state-of-the-art methods. SimSeg† [49] is trained with a subset of COCO Stuff in their paper. For a fair comparison, we reproduce their method on the full COCO Stuff with their officially released code. RN101: ResNet-101 [15]; EN-B7: EfficientNet-B7 [41]. ADE, PC, and VOC denote ADE20K [54], Pascal Context [35], and Pascal VOC [10], respectively.

Method	VL-Model	Training Dataset	ADE-150	ADE-847	PC-59	PC-459	VOC
Group-VIT [47]	rand. init.	CC12M+YFCC	-	-	22.4	-	52.3
LSeg+ [22]	ALIGN RN101	COCO	13.0	2.5	36.0	5.2	59.0
OpenSeg [12]	ALIGN RN101	COCO	15.3	4.0	36.9	6.5	60.0
LSeg+ [22]	ALIGN EN-B7	COCO	18.0	3.8	46.5	7.8	-
OpenSeg [12]	ALIGN EN-B7	COCO	21.1	6.3	42.1	9.0	-
OpenSeg [12]	ALIGN EN-B7	COCO+Loc. Narr.	28.6	8.8	48.2	12.2	72.2
SimSeg [49]	CLIP ViT-B/16	COCO	20.5	7.0	47.7	8.7	88.4
SimSeg† [49]	CLIP ViT-B/16	COCO	21.1	6.9	51.9	9.7	91.8
OVSeg [27]	CLIP ViT-B/16	COCO	24.8	7.1	53.3	11.0	92.6
MAFT [21]	CLIP ViT-B/16	COCO	29.1	10.1	53.5	12.8	90.0
SAN [50]	CLIP ViT-B/16	COCO	27.5	10.1	53.8	12.6	94.0
SCAN (Ours)	CLIP ViT-B/16	COCO	**30.8**	**10.8**	**58.4**	**13.2**	**97.0**
MaskCLIP [9]	CLIP ViT-L/14	COCO	23.7	8.2	45.9	10.0	-
SimSeg† [49]	CLIP ViT-L/14	COCO	21.7	7.1	52.2	10.2	92.3
OVSeg [27]	CLIP ViT-L/14	COCO	29.6	9.0	55.7	12.4	94.5
ODISE [48]	CLIP ViT-L/14	COCO	29.9	11.1	57.3	14.5	-
SAN [50]	CLIP ViT-L/14	COCO	32.1	12.4	57.7	15.7	94.6
SCAN (Ours)	CLIP ViT-L/14	COCO	**33.5**	**14.0**	**59.3**	**16.7**	**97.2**

this dataset. ADE20K-150 contains 150 semantic classes whereas ADE20K-847 has 847 classes. The images of both are the same. Pascal Context is an extension of Pascal VOC 2010, containing 4,998 training images and 5,005 validation images. We take the commonly used PC-59 and challenging PC-459 version for validation. Pascal VOC contains 11,185 training images and 1,449 validation images from 20 classes. We use the provided augmented annotations.

Evaluation Metric Following previous works [8, 12, 49], we take the *mean-intersection-over-union* (mIoU) as the metric to compare our model with previous state-of-the-art methods. In addition, we also report the corresponding results measured by our proposed SG-IoU.

4.2. Implementation Details

For segmentation model, our implementation is based on `detectron2` [44]. All image-based models are trained with batch size of 32 and training iteration of 120k. The base learning rate is 0.00006 with a polynomial schedule. The shortest edge of input image is resized to 640. For data augmentation, random flip and color augmentation are adopted. The weight decay of the segmentation model is 0.01. The backbone of segmentation model is Swin Transformer-Base [32]. The CLIP [38] version is ViT-L/14, implemented by OpenCLIP. For the weights of the loss function, we set 5 and 2 for segmentation loss and classification loss, respectively. The segmentation loss consists of dice loss and cross entropy loss. The classification loss is cross entropy loss. Other hyperparameters are the same as Mask2Former [7]. For fine-tuned CLIP, the training process is the same as OVSeg [27].

4.3. Main Results

We compare our model with existing state-of-the-art approaches in Table 1. To make it clear, we group the methods according to the utilized vision-language model and report the performance of our SCAN with CLIP ViT-B/16 as well as ViT-L/14 [38]. It can be seen that with global distribution prior, our model achieves the best performance on all popular benchmarks under both ViT-B and ViT-L. With ViT-B/16, our model reaches 30.8 and 58.4 on ADE-150 and PC-59, surpassing previous methods by a large margin. For ViT-L/14, our SCAN overpasses previous state-of-the-art by about 1.5% on ADE-150 and ADE-847. On PC-59 and PC-459, SCAN achieves 59.3 and 16.7, respectively.

4.4. Evaluation with SG-IoU

The above comparisons are based on vanilla evaluation system. As explained in Section 3.3, there exists problems when directly using traditional mIoU as evaluation metric for open-vocabulary segmentation performance under existing datasets. Therefore, we also report the results evaluated with the proposed SG-IoU in Table 2. By taking semantic relationships between different categories into account, the performance would improve and the gap between various methods is also different from Table 1. More analysis please see the supplementary materials.

Table 2. Evaluation with SG-IoU as metric of our SCAN and some open-source methods. For the sake of comparison, we report the results of vanilla mIoU in gray color.

Method	ADE-150	ADE-847	PC-459
SimSeg[49]	22.6 / 20.5	8.1 / 7.0	9.3 / 8.7
OVSeg [27]	30.5 / 29.6	9.5 / 9.0	12.7 / 12.4
MAFT [21]	30.3 / 29.1	11.5 / 10.1	13.4 / 12.8
SAN [50]	33.7 / 32.1	13.2 / 12.4	16.2 / 15.7
SCAN(Ours)	**34.2** / 33.5	**14.6** / 14.0	**17.2** / 16.7

Table 3. Ablation experiments on the proposed modules.

Method	ADE-150	ADE-847	PC-59
Baseline	31.5	11.4	57.1
+ SIM	32.9	13.3	58.6
+ CS	32.8	12.6	58.3
+ Both	**33.5**	**14.0**	**59.3**

Table 4. Different structure design of Semantic Integration Module(SIM). F_{CLS}^{final}, F_{HW} and LFE denote leveraging the final [CLS] embedding, selected spatial token embeddings, and low-frequency enhancement strategy, respectively.

F_{CLS}^{final}	F_{HW}	LFE	ADE-150	ADE-847	PC-59
			31.5	11.4	57.1
✓			32.5	12.7	57.5
	✓		32.0	11.9	57.5
✓	✓		32.6	13.0	58.1
	✓	✓	32.4	12.2	57.7
✓	✓	✓	**32.9**	**13.3**	**58.6**

4.5. Ablation Study

Model Component Analysis Table 3 shows the ablation study about the proposed Semantic Integration Module (SIM) and Contextual Shift (CS) strategy. It can be seen that when using SIM or CS alone, the model performance improves because of the injection of CLIP global semantic prior. When using both of them, the model achieves the best performance on all benchmarks.

Analysis on Semantic Integration Module Table 4 reports the influence of different structure design of SIM. To verify the effectiveness precisely, corresponding experiments are conducted without contextual shift strategy. It can be seen that without any semantic guidance, the model performs poorly. Since the final [CLS] embedding contains rich semantic information and has been aligned with textual domain, incorporating it to proposal embeddings can greatly prevent overfitting to training categories and improve final performance. Besides, the spatial tokens from middle layer of CLIP also contain local semantic perceptions and contribute to open-vocabulary segmentation. But due to the potential high-frequency noise, the employed spectral enhancement operation helps to fully utilize such fine-grained semantic source. SIM achieves the best performance when using all of them.

Table 5. Selected CLIP layers and frequency kernels for SIM.

Layers	Kernel	ADE-150	ADE-847	PC-59
15, 18, 21	7, 5, 3	32.5	12.9	58.3
15, 18, 21	9, 7, 3	32.4	13.0	58.5
12, 18, 24	7, 5, 3	33.3	13.6	58.9
12, 18, 24	9, 7, 3	**33.5**	**14.0**	**59.3**
12, 18, 24	11, 9, 7	33.0	13.4	58.4

Table 6. Comparison of different substitution sources.

Method	ADE-150	ADE-847	PC-59
None	32.7	13.3	58.6
Random noise	32.3	13.0	57.8
Original background	32.3	13.1	58.0
Learnable prompt	33.1	13.2	58.8
SCAN(Ours)	**33.5**	**14.0**	**59.3**

We also conduct experiments about the selected layer of spatial token embeddings from original CLIP and the selection of frequency kernels, *i.e.*, the cutoff frequency σ. The results are shown in Table 5. We can see that the spatial embeddings should hold information of multi-level granularity rather than all from deep layers. We attribute this to the complementary nature of the multi-granularity information. Besides, an appropriate cut-off frequency is also required due to the presence of undesirable texture noise.

Contextual Shift Strategy Here we compare different source of substitution tokens for contextual shift strategy as well as the influence of related location and ratios.

Sources of substitutions: to relieve the problems of domain bias, we randomly replace patch embeddings belong to background area with [CLS] embedding from corresponding layers of original CLIP with natural image as input. We also experiment with other replacing strategies under the same replacement percentage in Table 6. Specifically, we trial with utilizing random noise to replace (Random noise), randomly preserving the original pixels (Original background), and with learnable tokens (Learnable prompt) [27]. Results show that although such strategies can also disrupt the background shape, they struggles to simultaneously maintain contextual semantics and mitigate domain bias, leading to performance degradation. Besides, the learnable prompt is concerned about overfitting to seen semantics. The mIoU of taking learnable prompt is increased on ADE-150 and PC-59 datasets, which are similar to training space. But the performance drops on the more challenging benchmark ADE-847.

Location and ratios of substitutions: Table 7 presents the results of replacing background patches at different layers of CLIP with different ratios. It can be seen that the replacing should occur at shallow layers. If the contextual shift occurs after 11-th layer, the performance would drops. We analyze this is because the biased background region is already sensed by the shallow layers of CLIP and brings er-

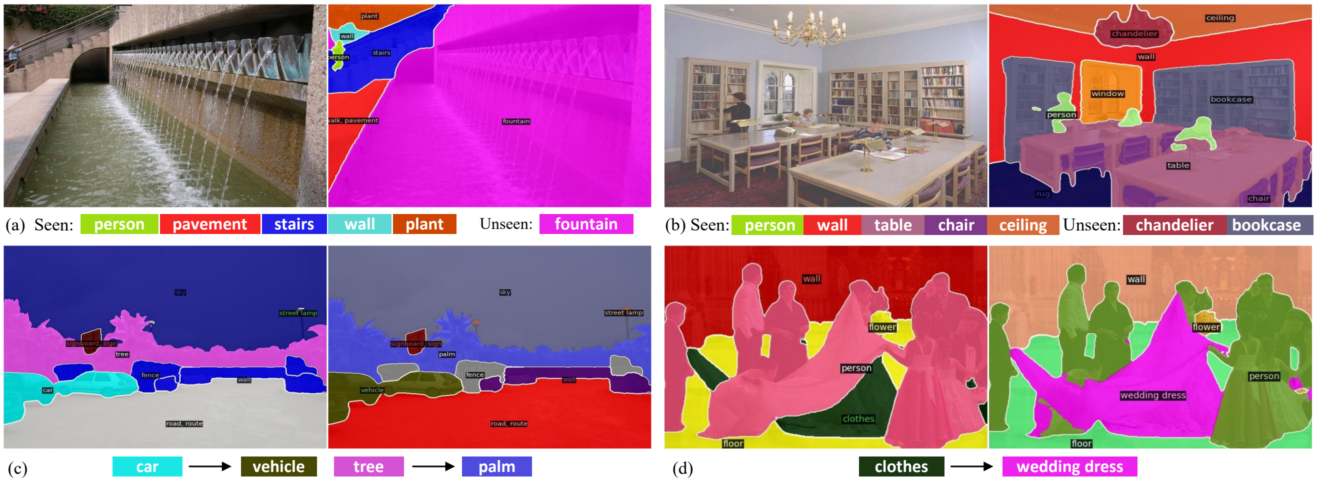


Figure 6. Visualization segmentation results. (a) and (b) demonstrate the excellent segmentation of our method for seen and unseen categories. (c) and (d) displays the adaptability for flexible text query. Best viewed in color.

Table 7. Replacement ratios and layers within CS strategy.

Layers	Ratios	ADE-150	ADE-847	PC-59
W/o Contexutal Shift	-	32.9	13.3	58.6
1, 3, 5, 7, 9	10%	**33.5**	13.6	59.4
1, 3, 5, 7, 9	20%	33.1	13.8	**59.5**
1, 3, 5, 7, 9	30%	**33.5**	**14.0**	59.3
1, 3, 5, 7, 9	40%	32.9	13.6	59.3
1, 3, 5, 7	30%	32.8	13.6	59.0
1, 2, 3, 4, 5	30%	32.2	13.2	58.9
11, 13, 15, 17, 19	30%	32.2	12.0	59.1

Table 8. Improvement of model proposal embedding and domain-biased CLIP classification, respectively.

	ADE-150	ADE-847	PC-59
(a) Baseline			
Only Mask Embedding	24.8	9.3	56.7
Only CLIP Embedding	27.7	10.0	48.2
Both	31.5	11.4	57.1
(b) Our SCAN			
Only Mask Embedding	26.3↑ (**1.5**)	11.1↑ (**1.8**)	57.9↑ (**1.2**)
Only CLIP Embedding	28.9↑ (**1.2**)	10.4↑ (**0.4**)	49.9↑ (**1.7**)
Both	33.5↑ (**2.0**)	14.0↑ (**2.6**)	59.3↑ (**2.2**)

roneous prior. For replacing ratios, we find that excessively high proportion of replacement would make global context impair local area judgements, as shown in Layers of (1, 3, 5, 7, 9) with 40% replacement ratios. The results of layers (1, 2, 3, 4, 5) with 30% replacement ratios also drops on ADE-150 [54]. We posit that this can be attributed to the fact that successive replacement tends to induce overly global representation effects, which are akin to high replacement ratios.

Improvement of In-Vocabulary and Domain-Biased Embedding To prove the proposed SIM and CS strategy can relieve the overfitting problem of proposal embeddings and image domain bias for CLIP, we test the performance gains of using them alone. From Table 8 we can see that by calibrating corresponding space, the cross-modal alignment of both mask embedding and CLIP embedding have been remarkably improved. With both calibration on mask embedding and CLIP embedding, performance is more significantly improved.

4.6. Visualization

Figure 6 shows some segmentation cases of our SCAN. It can be seen that our method achieves excellent segmentation performance on various scenarios. Specifically, (a) and (b) demonstrate the excellent segmentation of our method for seen and unseen categories. (c) and (d) display the adaptability for flexible text query, *e.g.*, change “tree” to “palm” and “clothes” to “wedding dress”.

5. Conclusion

We present a Semantic-assisted CAlibration Network (SCAN) in this paper to boost vision-language alignment performance. In SCAN, SIM is proposed to calibrate the mask proposal embedding and relieve the overfitting problem. To compensate global context and mitigate the image domain bias, CS strategy is adopted for CLIP prediction. Extensive experiments show that SCAN achieves state-of-the-art performance on all popular open-vocabulary segmentation benchmarks. Besides, we focus on the problem of existing evaluation system that neglects relationships across classes, and propose a new metric called SG-IoU.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62206153 and NO. 62322608, Shenzhen Key Laboratory of Ubiquitous Data Enabling (Grant No. ZDSYS20220527171406015).

References

[1] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *NeurIPS*, 2019. 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. *TPAMI*, 2018. 1

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[4] Xinlei Chen, Aveek Purohit, Carlos Ruiz Dominguez, Stefano Carpin, and Pei Zhang. Drunkwalk: Collaborative and adaptive planning for navigation of micro-aerial sensor swarms. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, pages 295–308, 2015. 1

[5] Xuecheng Chen, Haoyang Wang, Zuxin Li, Wenbo Ding, Fan Dang, Chengye Wu, and Xinlei Chen. Deliversense: Efficient delivery drone scheduling for crowdsensing with deep reinforcement learning. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pages 403–408, 2022. 1

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 3

[7] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 6

[8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 1, 2, 3, 5, 6

[9] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 6

[10] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015. 5, 6

[11] Yan Fang, Feng Zhu, Bowen Cheng, Luoqi Liu, Yunchao Wei, and Yao Zhao. Locating noise is halfway denoising for semi-supervised segmentatio. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1

[12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 1, 2, 6

[13] Meng-Hao Guo, Chengze Lu, Qibin Hou, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022. 1

[14] 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, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[16] Shuting He, Henghui Ding, and Wei Jiang. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In *CVPR*, 2023. 1

[17] Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. Segment and caption anything. *arXiv preprint arXiv:2312.00869*, 2023. 1

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3

[19] 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 *ICML*, pages 4904–4916, 2021. 1

[20] Zhuozhu Jian, Zejia Liu, Haoyu Shao, Xueqian Wang, Xinlei Chen, and Bin Liang. Path generation for wheeled robots autonomous navigation on vegetated terrain. *IEEE Robotics and Automation Letters*, 2023. 1

[21] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *arXiv preprint arXiv:2310.00240*, 2023. 5, 6, 7

[22] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 1, 3, 6

[23] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 3

[24] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. *NeurIPS*, pages 35313–35325, 2022. 3

[25] 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. 3

[26] Zuxin Li, Fanhang Man, Xuecheng Chen, Baining Zhao, Chenye Wu, and Xinlei Chen. Tract: Towards large-scale crowdsensing with high-efficiency swarm path planning. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pages 409–414, 2022. 1

[27] 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. *arXiv preprint arXiv:2210.04150*, 2022. 2, 5, 6, 7

[28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 5

[29] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *ECCV*, pages 648–665, 2022. 1

[30] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *ECCV*, pages 468–486, 2022. 1

[31] Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. *arXiv preprint arXiv:2312.01623*, 2023. 1

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 6

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3

[34] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *arXiv preprint arXiv:2305.17011*, 2023. 2

[35] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 5, 6

[36] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. Siri: A simple selective retraining mechanism for transformer-based visual grounding. In *ECCV*, 2022. 3

[37] Mengxue Qu, Yu Wu, Yunchao Wei, Wu Liu, Xiaodan Liang, and Yao Zhao. Learning to segment every referring object point by point. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[38] 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 *ICML*, 2021. 1, 2, 3, 4, 5, 6

[39] Jiyuan Ren, Yanggang Xu, Zuxin Li, Chaopeng Hong, Xiao-Ping Zhang, and Xinlei Chen. Scheduling uav swarm with attention-based graph reinforcement learning for ground-to-air heterogeneous data communication. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pages 670–675, 2023. 1

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[41] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 6

[42] Haoyang Wang, Xuecheng Chen, Yuhan Cheng, Chenye Wu, Fan Dang, and Xinlei Chen. H-swarmloc: Efficient scheduling for localization of heterogeneous mav swarm with deep reinforcement learning. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 1148–1154, 2022. 1

[43] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 3

[44] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[45] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. In *CVPR*, 2019. 1, 2

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NIPS*, 2021. 1

[47] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 6

[48] 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 *CVPR*, pages 2955–2966, 2023. 1, 2, 6

[49] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 1, 2, 3, 5, 6, 7

[50] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 2, 5, 6, 7

[51] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *ICONIP*, pages 264–274, 2019. 3

[52] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *NeurIPS*, 2019. 3

[53] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *ICCV*, 2021. 1

[54] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 2, 5, 6, 8