Side Adapter Network for Open-Vocabulary Semantic Segmentation

Mengde Xu[*1], Zheng Zhang[*1,2], Fangyun Wei[2], Han Hu[2], Xiang Bai[1]

[1]HUST,[2]Microsoft

CVPR 2023 Highlight

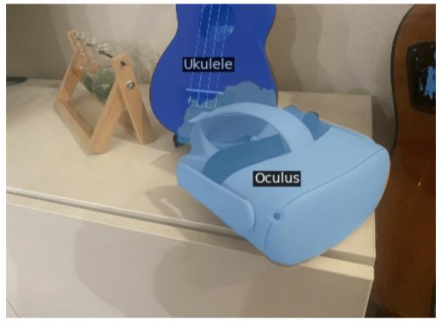
Background

1. Open-Vocabulary Seg

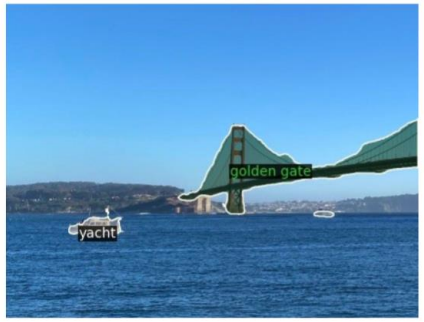
Def. Recognizing and segmenting the visual elements of any category.

Vision-language models, e.g., CLIP, learn rich multi-modal features from billion-scale image-text pairs. Witnessing its superior **open-vocabulary classification ability**, prior works propose to use pre-trained vision-language models for open-vocabulary segmentation.

Query: saturn V, blossom



Query: Oculus, Ukulele



Query: golden gate, yacht

Background

1. Open-Vocabulary Seg

Benckmark: Performing zero-shot segmentation on arbitrary datasets without dataset-specific adaption.

Table 1. The mIoU results of open-vocabulary generalist models and supervised specialist models. Results for SPNet and ZS3Net on PAS-20 are reported from [23]. Results for ZegFormer on PAS-20 are recalculated by us. SimBaseline [40], ZegFormer [11] and OpenSeg [16] are using the same COCO images, *i.e.*, the 2017 splits with 118K images, but with different annotations. COCO-Stuff-156/171 denotes using COCO Stuff mask annotations of 156/171 categories. Under the R101c model scale, our model significantly outperforms other open-vocabulary models. Our largest Swin-Base model can match the performance of some supervised specialist models in 2017.

method	backbone	training dataset	A-847	PC-459	A-150	PC-59	PAS-20
Open-vocabulary generalist models							
SPNet [37]	R-101	PASCAL-15	-	-	-	24.3	18.3
ZS3Net [4]	R-101	PASCAL-15	-	-	-	19.4	38.3
LSeg [23]	R-101	PASCAL-15	-	-	-	-	47.4
LSeg+ [16]	R-101	COCO Panoptic	2.5	5.2	13.0	36.0	59.0
SimBaseline [40]	R-101c	COCO-Stuff-156	-	-	15.3	-	74.5
ZegFormer [11]	R-50	COCO-Stuff-156	-	-	16.4	-	80.7
OpenSeg [16]	R-101	COCO Panoptic	4.0	6.5	15.3	36.9	60.0
OVSeg (Ours)	R-101c	COCO-Stuff-156	7.0	10.4	24.0	51.7	89.2
OVSeg (Ours)	R-101c	COCO-Stuff-171	**7.1**	**11.0**	**24.8**	**53.3**	**92.6**
LSeg+ [16]	Eff-B7	COCO Panoptic	3.8	7.8	18.0	46.5	-
OpenSeg [16]	Eff-B7	COCO Panoptic	6.3	9.0	21.1	42.1	-
OVSeg (Ours)	Swin-B	COCO-Stuff-171	**9.0**	**12.4**	**29.6**	**55.7**	**94.5**

Background

2. Previous method:

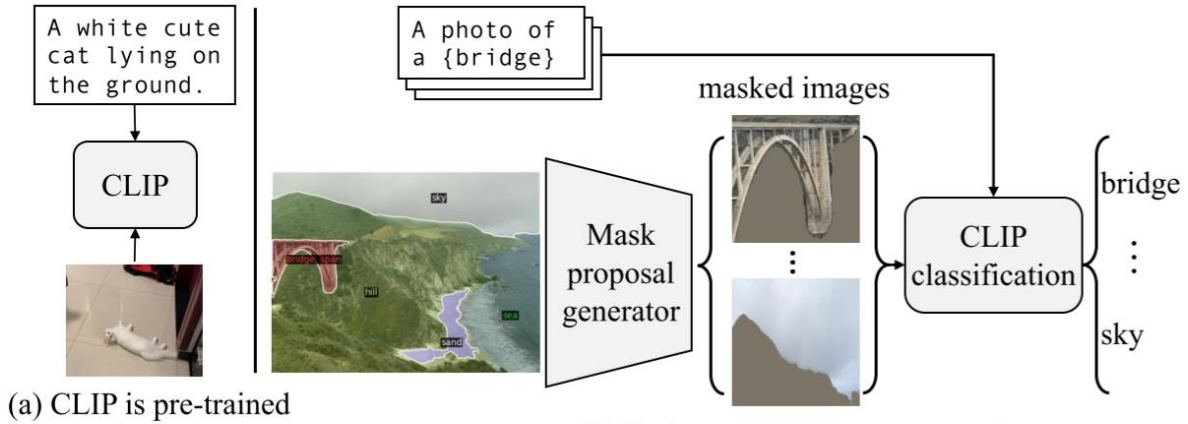
First: generate class-agnostic mask proposals then

Second: leverage pre-trained CLIP to perform open-vocabulary classification.

Their success relies on two **assumptions**:

(1) the model can generate class-agnostic mask proposals

(2) pre-trained CLIP can transfer its classification performance to masked image proposals.



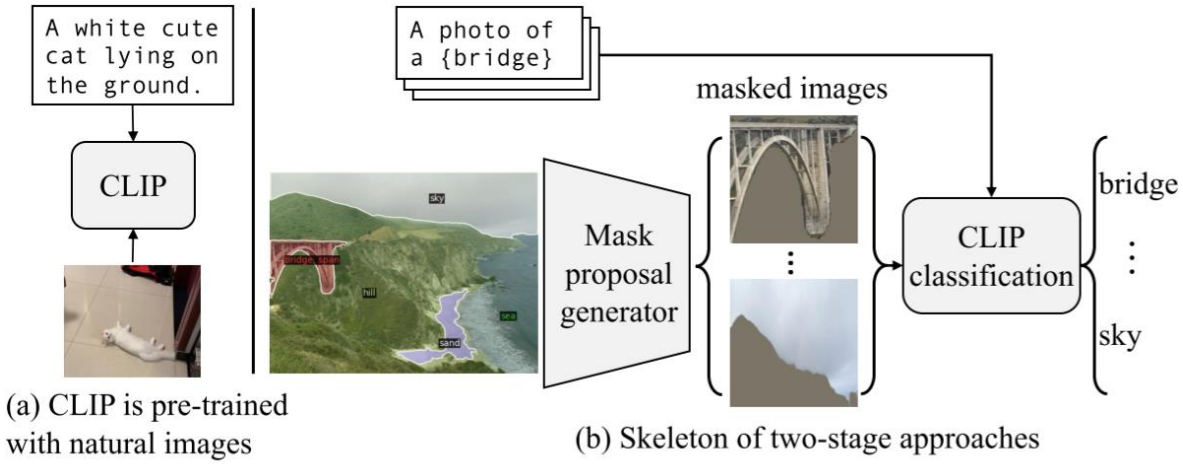
(a) CLIP is pre-trained with natural images

(b) Skeleton of two-stage approaches

Background

2. Previous method:

（two-stage）

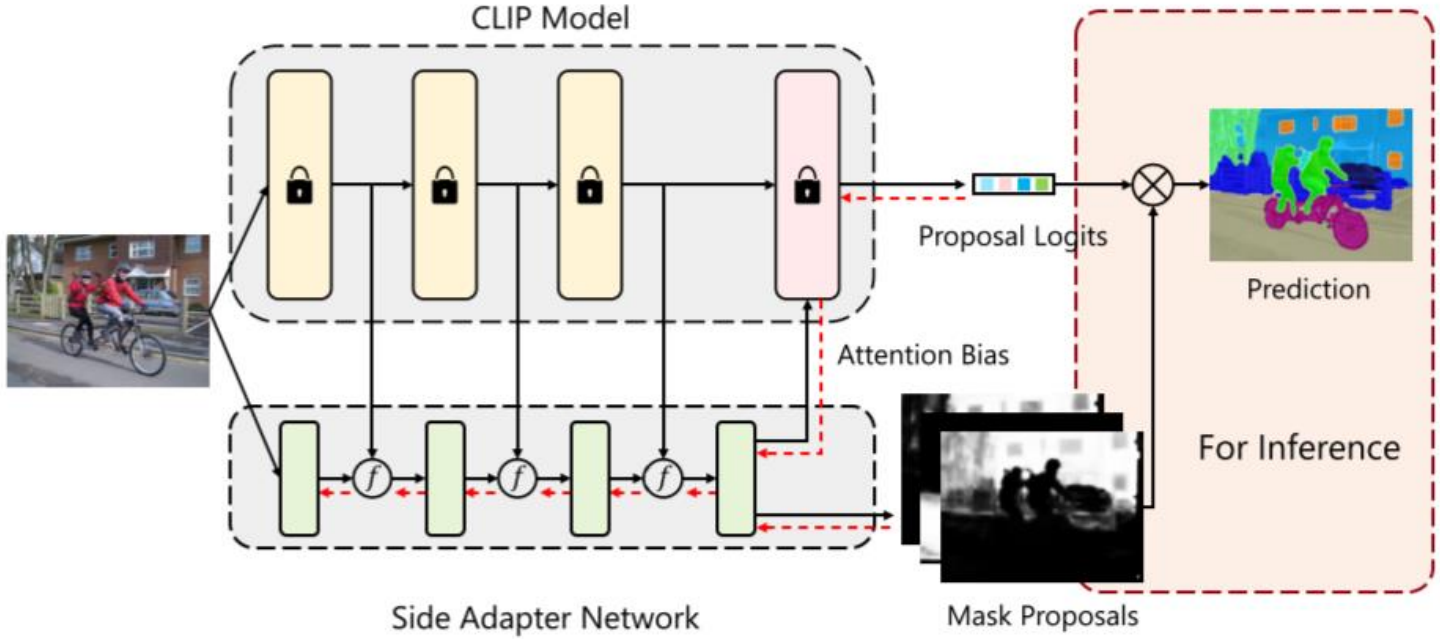


Problem:

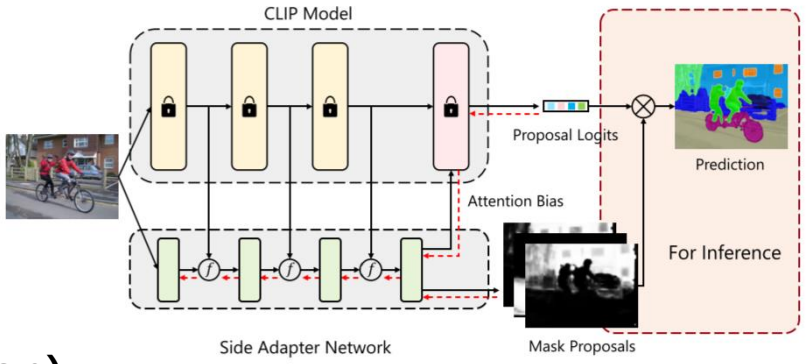
1. CLIP的feature无法被 mask proposal generator 利用
2. 由于CLIP只能对image进行预测，导致得到了mask之后还需要去crop/mask原图，再把这些处理过的images送到CLIP，这样可能就需要反复forward很多次，效率比较低。
3. 同时，也因为crop和mask，存在输入偏移的问题，使CLIP分类能力下降。
4. Mask proposal的质量直接影响分类的结果

Model

Our approach is an **end-to-end** framework, the mask prediction is **lightweight** and **CLIP-aware**, and the mask recognition is **decoupled** from mask prediction.

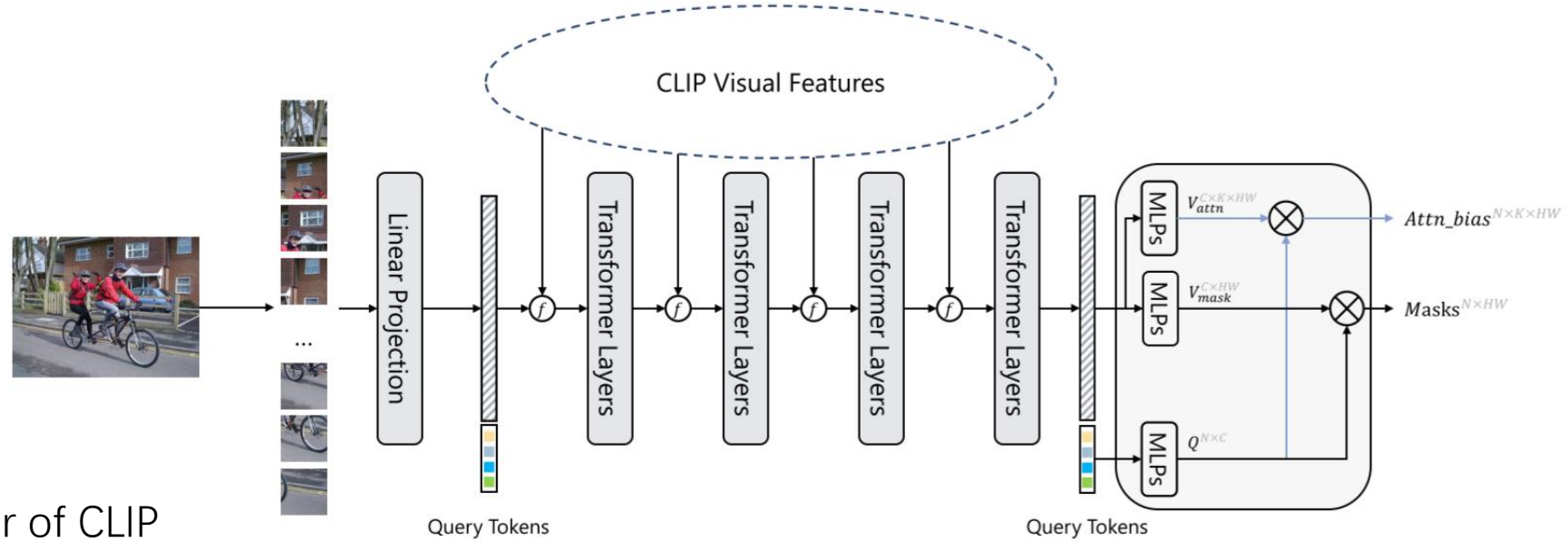


Key idea



1. Feature fusion on visual tokens

(leveraging the CLIP visual features & CLIP-aware mask prediction)

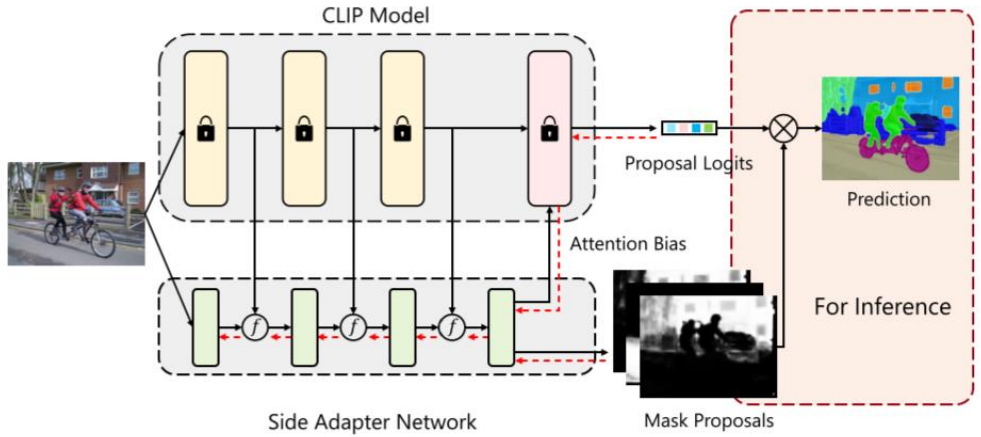


{stem, 3, 6, 9} layer of CLIP
{stem, 1, 2, 3} layer of SAN.

Key idea

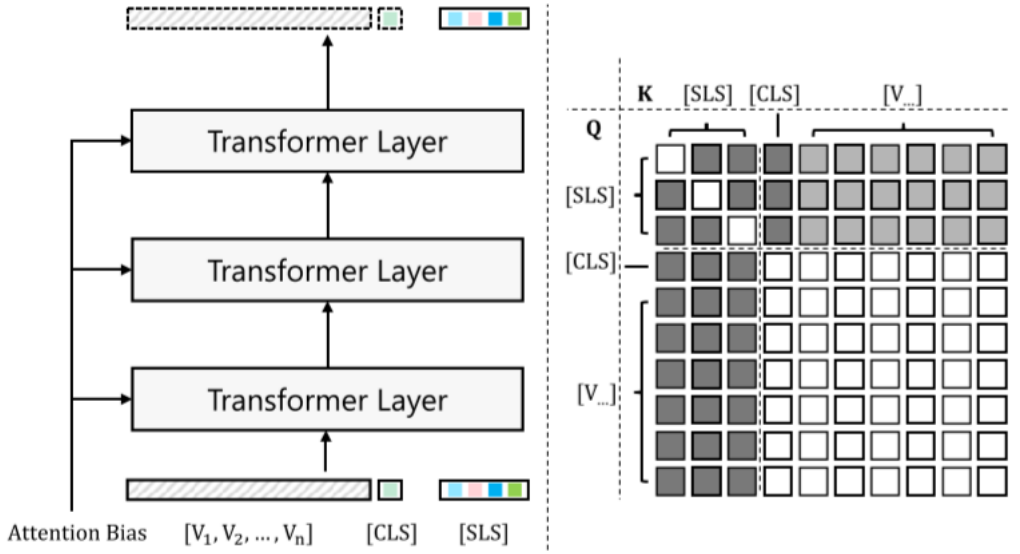
2. Mask recognition with attention bias

(CLIP for recognizing the class of mask proposals)



$$\mathbf{X}^{l+1}_{\texttt{[SLS]}} = \texttt{softmax}(\mathbf{Q}^{l}_{\texttt{[SLS]}}\mathbf{K}^{l}_{\texttt{visual}} + \mathbf{B}_k)\mathbf{V}^{l}_{\texttt{[SLS]}} \quad (3)$$

, where l indicates layer number, k indicates the k-th attention head, $\mathbf{Q}_{\texttt{[SLS]}} = \mathbf{W_q}\mathbf{X}_{\texttt{[SLS]}}$ and $\mathbf{V}_{\texttt{[SLS]}} = \mathbf{W_v}\mathbf{X}_{\texttt{[SLS]}}$ are query and value embedding of [SLS] tokens, and $\mathbf{K}_{\texttt{visual}} = \mathbf{W_k}\mathbf{X}_{\texttt{visual}}$ is the key embedding of visual tokens. $\mathbf{W_q}$, $\mathbf{W_k}$, $\mathbf{W_v}$ are weights of query, key, and value embedding layer, respectively.



Key idea

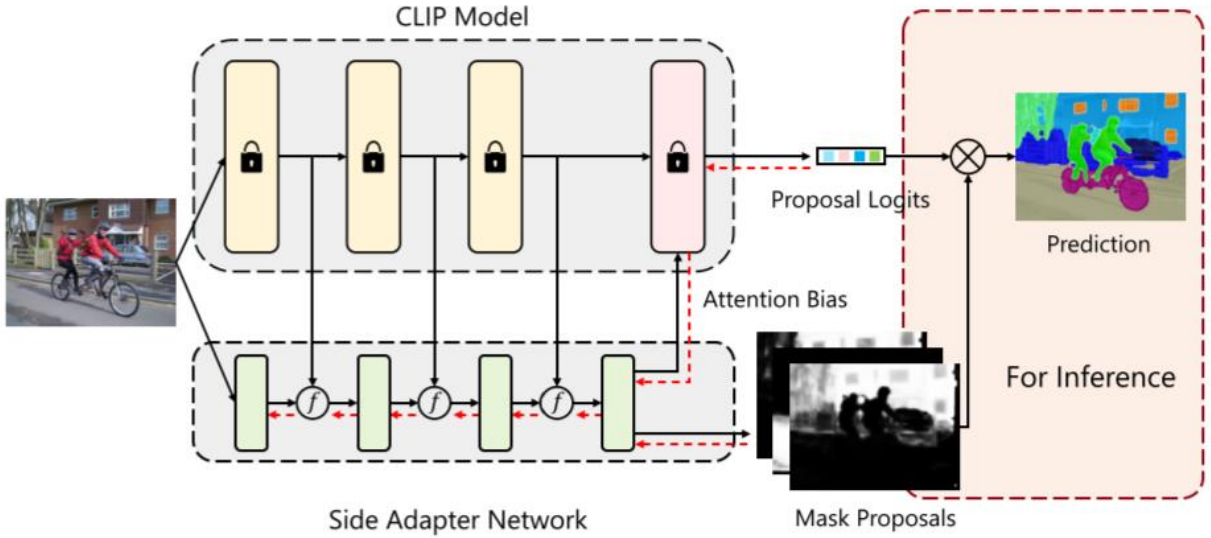
3. Segmentation map generation

(CLIP for recognizing the class of mask proposals)

$\mathbf{M} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times N}$ and the class prediction of masks $\mathbf{P} \in \mathbb{R}^{C \times N}$, we can compute the segmentation map:

$$\mathbf{S} = \mathbf{M} \times \mathbf{P}^{\mathrm{T}}$$

$$\mathbf{S} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$$



Experiments

1. Dataset and Evaluation Protocol

All models are trained on the training set of COCO Stuff and evaluated on other datasets.

Dataset	Label Sim. to COCO Stuff
Pascal VOC	0.91
Pascal Context-59	0.86
Pascal Context-459	0.70
ADE20K-150	0.73
ADE20K-847	0.57

Table 1. The label-set similarity between validation datasets and training set (*i.e.* COCO Stuff). Measured by Hausdorff distance and cosine similarity based on CLIP text encoder.

Experiments

2. System level comparison

Method	VL-Model	Training Dataset	ensemble.	ADE-847	PC-459	ADE-150	PC-59	VOC
Group-VIT [32]	rand. init.	CC12M+YFCC	no.	-	-	-	22.4	52.3
LSeg+ [12]	ALIGN RN101	COCO	no.	2.5	5.2	13.0	36.0	59.0
OpenSeg [12]	ALIGN RN101	COCO	no.	4.0	6.5	15.3	36.9	60.0
LSeg+ [12]	ALIGN EN-B7	COCO	no.	3.8	7.8	18.0	46.5	-
OpenSeg [12]	ALIGN EN-B7	COCO	no.	6.3	9.0	21.1	42.1	-
OpenSeg [12]	ALIGN EN-B7	COCO+Loc. Narr.	no.	8.8	12.2	28.6	48.2	72.2
SimSeg [33]	CLIP ViT-B/16	COCO	yes.	7.0	8.7	20.5	47.7	88.4
SimSeg†	CLIP ViT-B/16	COCO	yes.	6.9	9.7	21.1	51.9	91.8
OvSeg [22]	CLIP ViT-B/16	COCO	yes.	7.1	11.0	24.8	53.3	92.6
SAN(ours)	CLIP ViT-B/16	COCO	no.	**10.1** ± 0.23	**12.6** ± 0.44	**27.5** ± 0.34	**53.8** ± 0.57	**94.0** ± 0.21
SAN ensemble.	CLIP ViT-B/16	COCO	yes.	10.7 ± 0.22	13.7 ± 0.34	28.9 ± 0.42	55.4 ± 0.11	94.6 ± 0.11
MaskCLIP [10]	CLIP ViT-L/14	COCO	no.	8.2	10.0	23.7	45.9	-
SimSeg†	CLIP ViT-L/14	COCO	yes.	7.1	10.2	21.7	52.2	92.3
OvSeg [22]	CLIP ViT-L/14	COCO	yes.	9.0	12.4	29.6	55.7	94.5
SAN(ours)	CLIP ViT-L/14	COCO	no.	**12.4** ± 0.27	**15.7** ± 0.26	**32.1** ± 0.42	**57.7** ± 0.34	**94.6** ± 0.42
SAN ensemble.	CLIP ViT-L/14	COCO	yes.	13.7 ± 0.12	17.1 ± 0.18	33.3 ± 0.29	60.2 ± 0.31	95.5 ± 0.16

Table 2. Performance comparison with state-of-the-art methods. † SimSeg [33] 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 [14]; EN-B7: EfficientNet-B7 [29]; SAN ensemble. is the result using ensemble tricks, not the default setting.

Experiments

2. System level comparison

Method	Param. (M)	GFLOPs	FPS
SimSeg	61.1	1916.7	0.8
OvSeg*	147.2	1916.7	0.8
MaskCLIP*	63.1	307.8	4.1
SAN(ours)	**8.4**	**64.3**	**15.2**

Table 3. Training and testing efficiency comparison with other methods. *Param.* stands for the total number of trainable parameters in the methods in millions. The input image is of 640×640 resolution. And the clip model is ViT-B/16. * no official code available yet and we re-implement their methods following the description in their papers. OvSeg [22] has similar structures to SimSeg [33] but it finetuned the whole CLIP model,resulting in much more trainable parameters.

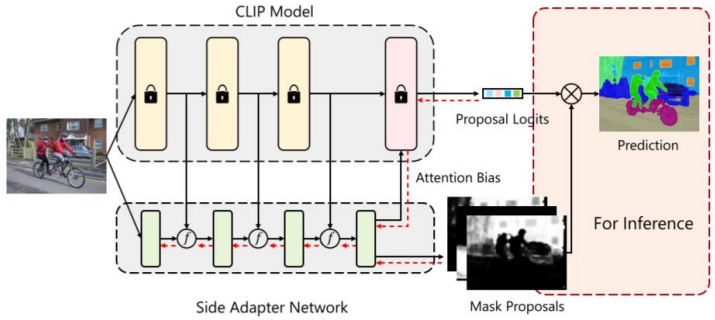
Experiments

3. Ablation Studies

(Importance of feature fusion.)

Description.	Layers	mIoU
w/o. fusion	none	21.1
single-fusion	stem	20.0
	3rd layer	24.1
	6th layer	26.2
	9th layer	27.1
multi-fusion	{6,9}-layers	27.0
	{3,6,9}-layers	27.7
	{stem,3,6,9}-layers	**27.8**

Table 4. Different feature fusion strategies. The last 3 layers of ViT-B/16 are used for mask prediction in all experiments.



#Feature Fusion Layers	#Recognition Layers	mIoU
12	12	27.6
11	1	25.9
10	2	27.3
9	3	**27.8**
6	6	26.9
3	9	23.8

Table 5. The trade-off between the number of feature fusion layers and the number of mask prediction layers. *Note*: the 2nd row (*i.e.* the {12,12} setting) is the *twice-forward* baseline.

Experiments

3. Ablation Studies

(Importance of CLIP-aware mask prediction.)

Description	Backbone	*CLIP-aware*	mIoU
SimSeg	ViT-B/16	no.	21.1
MaskCLIP	ViT-L/14	no.	23.7
two-stage training	ViT-B/16	no.	21.6
e2e training	ViT-B/16	yes.	**26.1** (+4.5)

Table 6. *Two-stage vs. end-to-end.* The significant improvement proves the importance of *CLIP-aware* mask prediction.

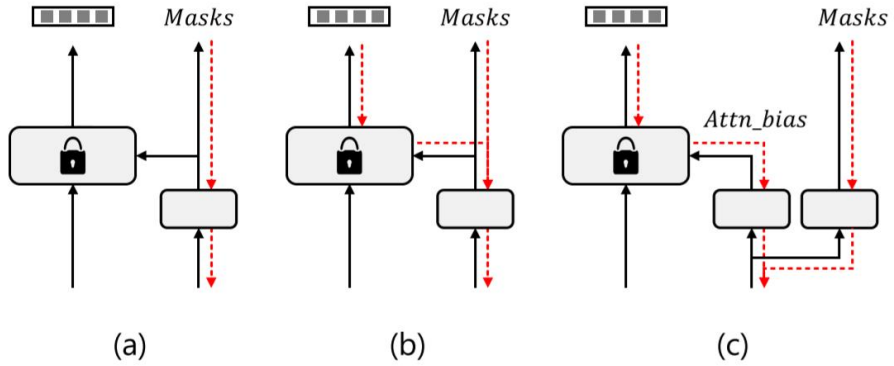


Figure 5. Design choice of mask prediction head. (a) Two-stage training with single head and blocking gradients from CLIP. (b) End-to-end training with single head (c) End-to-end training with decoupled head. The red dotted line indicates the gradient flow during training.

Experiments

3. Ablation Studies

(Asymmetric input resolution.)

Resolution.	GFLOPs	mIoU
192^2	39.4	25.3
224^2	44.3	26.3
320^2	64.3	**27.8**
448^2	106.3	26.1
640^2	213.4	24.6

Table 8. The influence of ViT-B/16 CLIP model input resolution. We vary CLIP input resolutions, while always using 640^2 images in the side-adapter network.

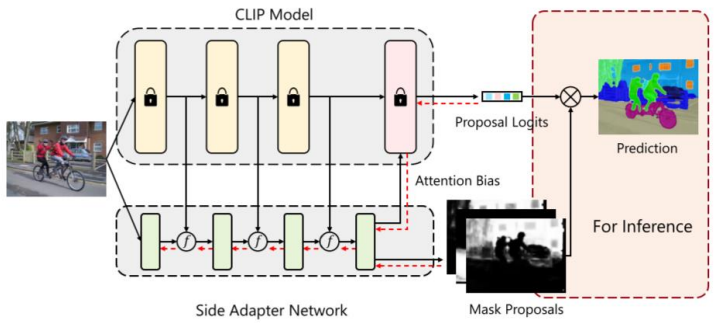
Description.	Resolution.	mIoU
fixed pos embed.	320^2	27.0
ft. pos embed.	320^2	**27.8**

Table 9. Fine-tuning the position embedding can improve the performance.

Experiments

3. Ablation Studies

(Discussion on the Parameter Efficiency)



Width of SAN	Param. (M)	GFLOPs	mIoU
144	4.2	53.6	26.7
192	6.1	58.6	27.4
240	8.4	64.3	**27.8**
288	11.1	70.9	27.3

Table 10. The influence of capacity of SAN. *Param.* stands for the total number of trainable parameters in the model in millions.


green lizard,lacerta viridis
water snake
terrapin
agama
jay
stingray
mud turtle
hen
goose
trilobite
tinca tinca,tench
brain coral
chlamydosaurus kingi,frilled lizard
toucan
vine snake
green mamba
grass snake,green snake
hunting spider,wolf spider
echidna,anteater,spiny anteater
black grouse
great white shark,white shark,carc
hummingbird
tarantula
tusker
peacock
alligator lizard
garden spider,aranea diademata
cock
hammerhead shark,hammerhead
hornbill