Open-Vocabulary Semantic Segmentation with Mask-adapted CLIP
--CVPR 2023

The University of Texas at Austin, Meta Reality Labs

https://jeff-liangf.github.io/projects/ovseg

目录

CONTENTS

01

PART

Introduction

demo presentation

class names

Golden gate, yacht,mountain

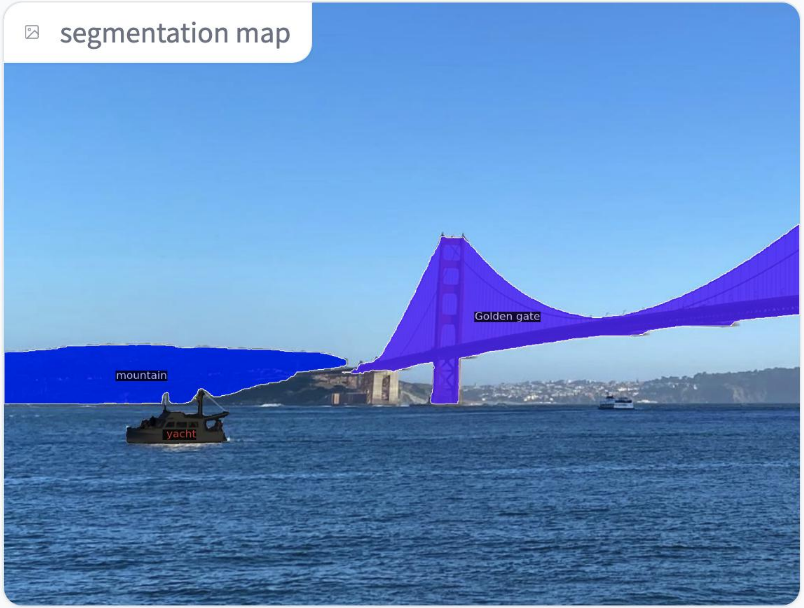
Proposal generator

Segment_Anything

MaskFormer

For Segment_Anything only, granularity of masks from 0 (most coarse) to 1 (most precise)

1



Open-vocabulary semantic segmentation

Train

person
bicycle
background

Test

motorbike

Input image

w/o zero shot

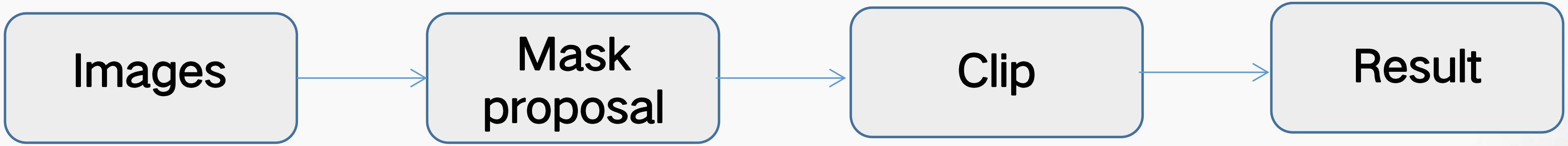
with zero shot

Related works （ECCV2022）

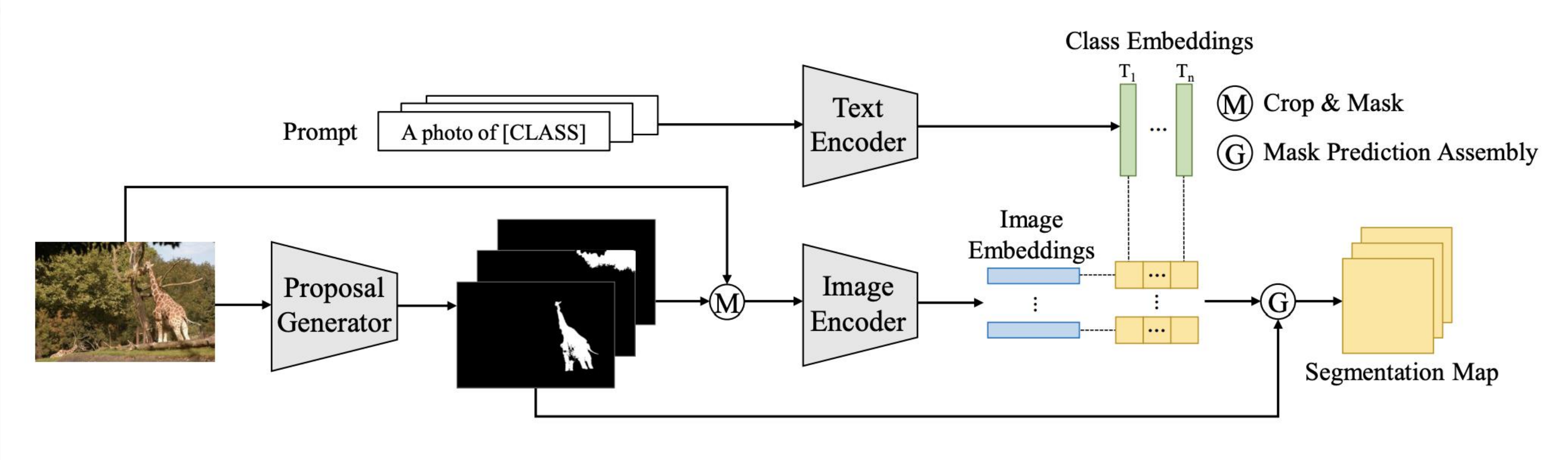
A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-language Model

Mengde Xu[1,3*], Zheng Zhang[1,3*], Fangyun Wei[3*], Yutong Lin[2,3], Yue Cao[3], Han Hu[3], and Xiang Bai[1†]

[1] Huazhong University of Science and Technology
[2] Xi'an Jiaotong University
[3] Microsoft Research Asia



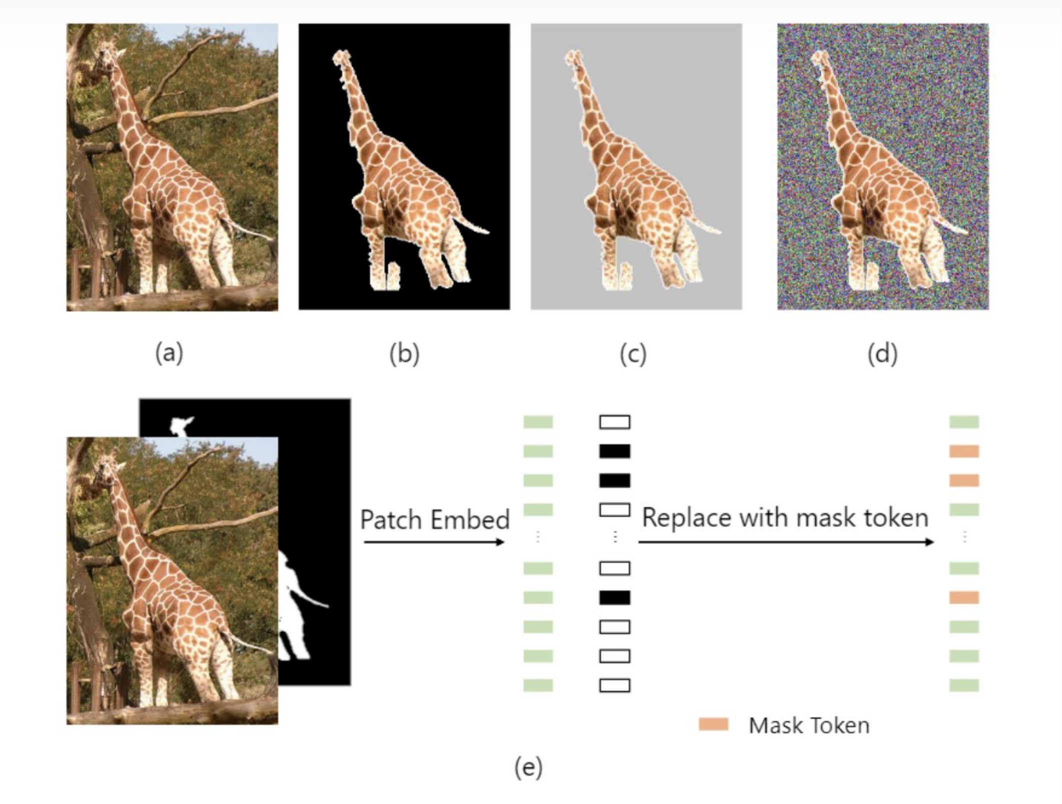
Related works



Segmantation model → Localization

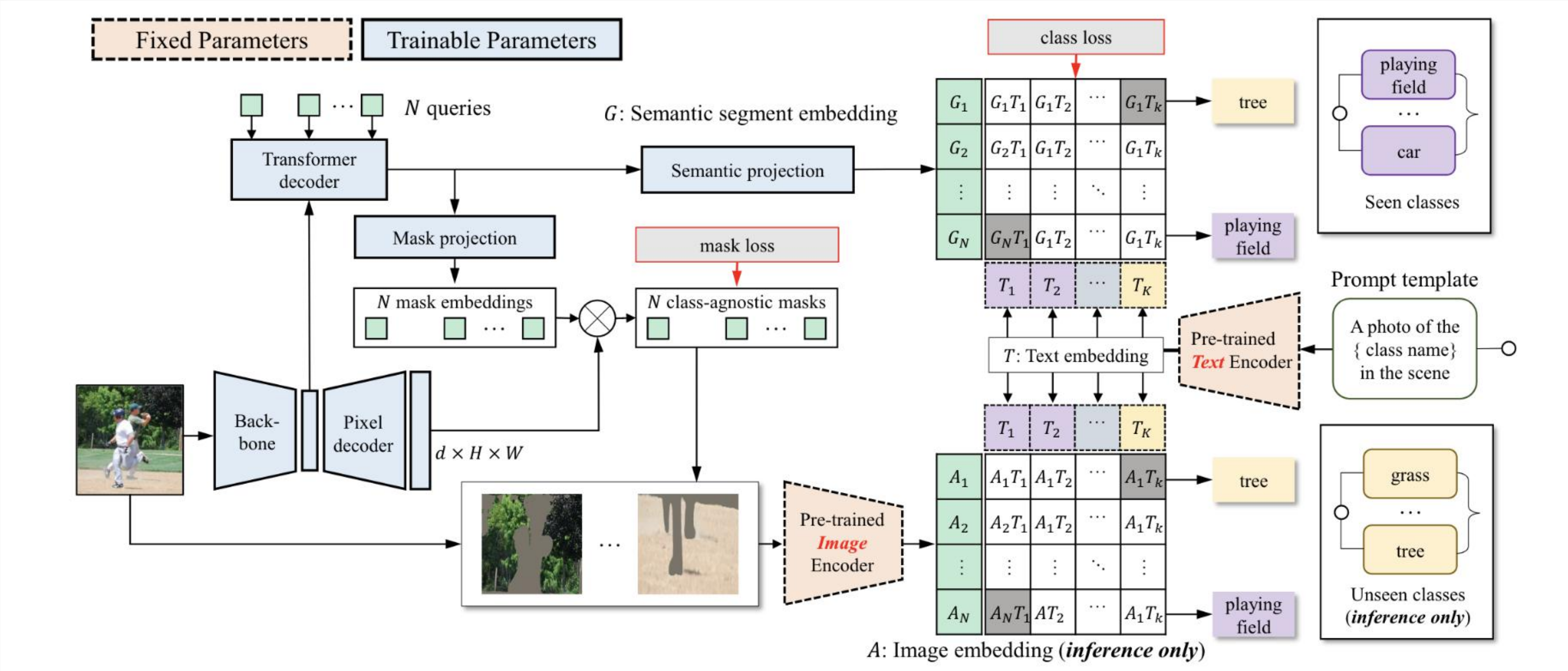
CLIP model → zero shot classification

Related works



Prompt	hIoU	mIoU	
		seen	unseen
Preserving	9.3	8.9	9.5
Zero	17.2	16.3	18.2
Mean Values	18.3	17.3	19.5
Pixel Prompts	Failed	-	-
Mask Token	Failed	-	-

Related works (Decoupling Zero-Shot Semantic Segmentation --CVPR 2022)

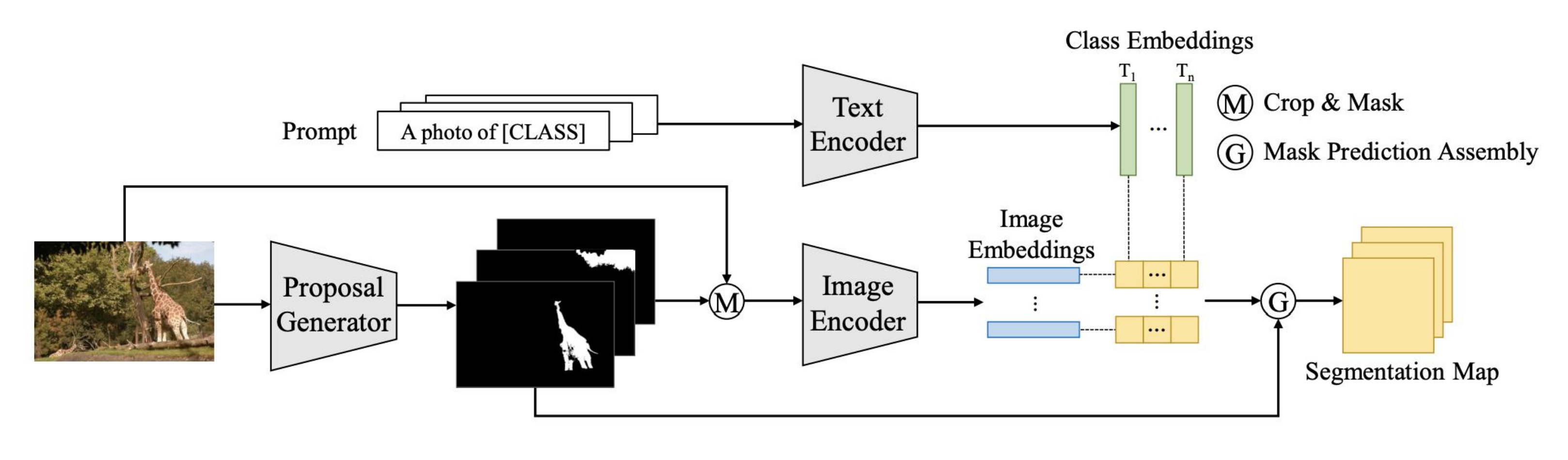


Related works

	preprocess	Seen	Unseen	Harmonic
ZegFormer-seg	-	**37.4**	21.4	27.2
ZegFormer	crop	36.6	19.7	25.6
	mask	36.0	31.0	33.3
	crop and mask	35.9	**33.1**	**34.4**

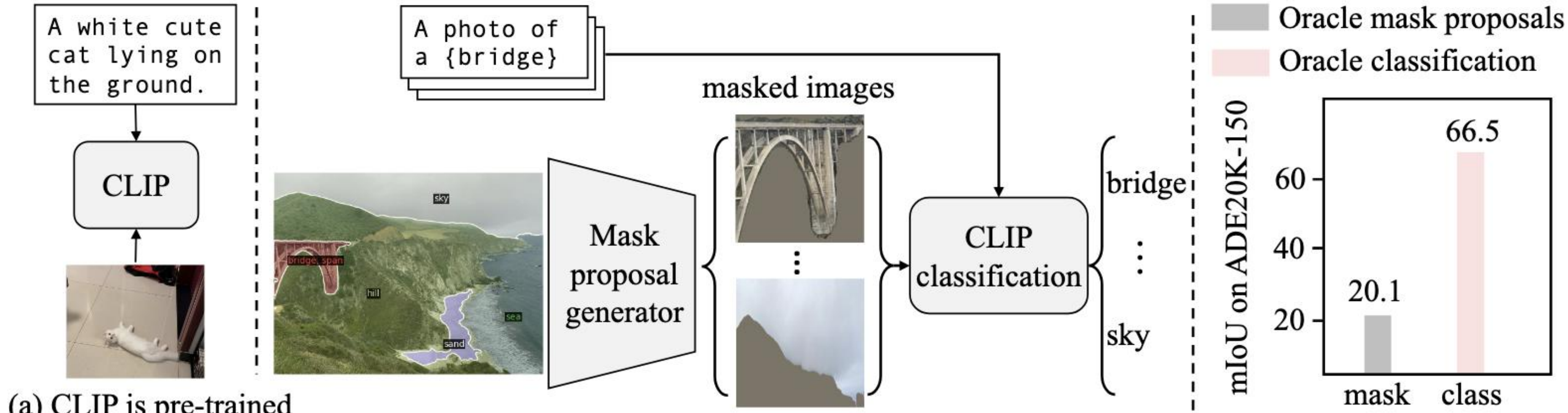
original image | crop | mask | mask and crop

Motivation



- model can generate class-agnostic mask proposals
- pre-trained CLIP can transfer its classification performance to masked image proposals.

Motivation



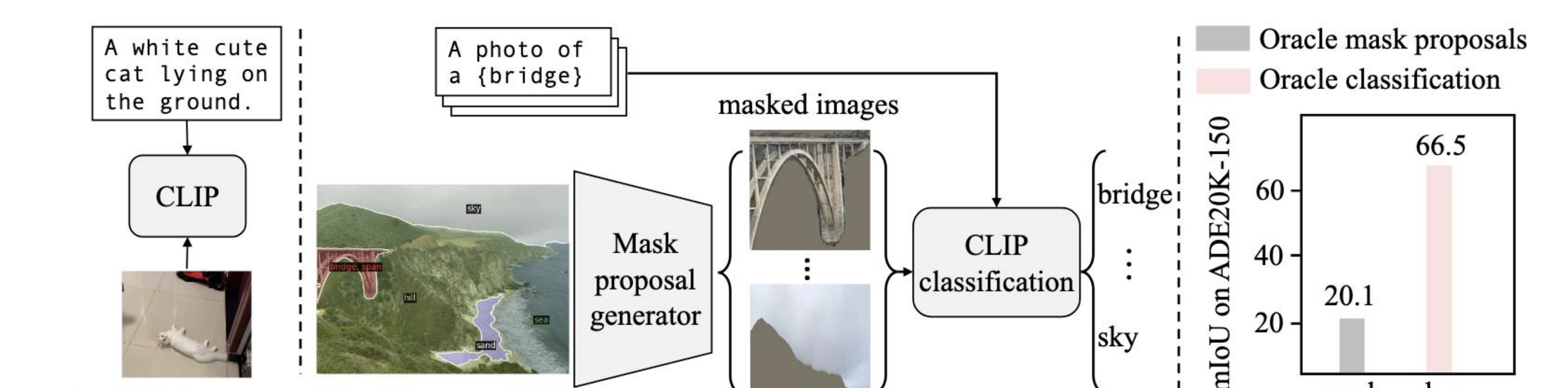
(a) CLIP is pre-trained with natural images

(b) Skeleton of two-stage approaches

(c) Bottleneck analysis

- Oracle masks + classification
- masks + Oracle classification

Motivation



(a) CLIP is pre-trained with natural images
(b) Skeleton of two-stage approaches
(c) Bottleneck analysis

- natural images <-> maksed proposals
- CLIP trained with minimal training augmentation

Contribution

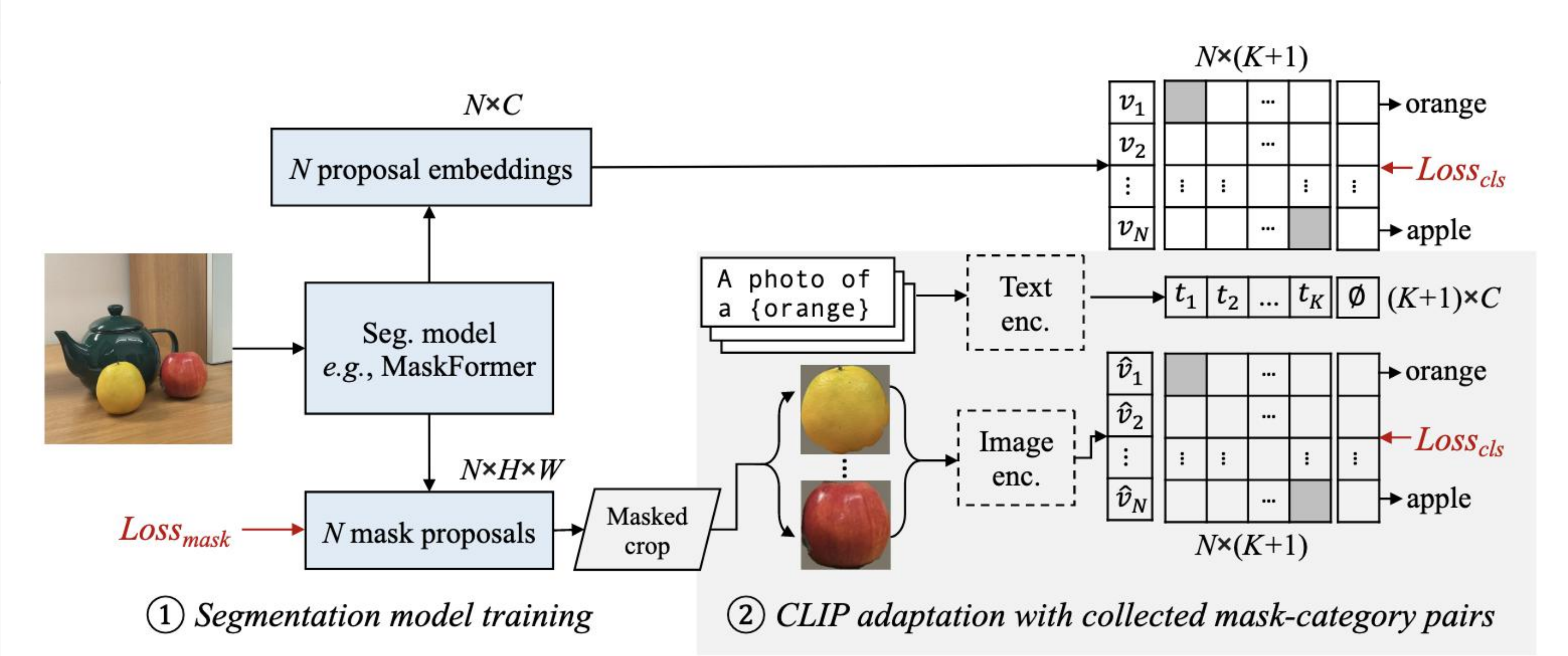
- Our analysis reveals the pre-trained CLIP does not perform well on mask proposals, making it the performance bottleneck of two-stage approaches.
- We collect diverse mask category pairs from captions to adapt CLIP for masked images and retain its generalization ability.
- We propose mask prompt tuning specifically for masked image adaptation. This method does not change CLIP's weight, enabling multi-task weight sharing.
- For the first time, we show open-vocabulary generalist models can match the performance of supervised specialist models in 2017 without dataset specific adaptations.

02

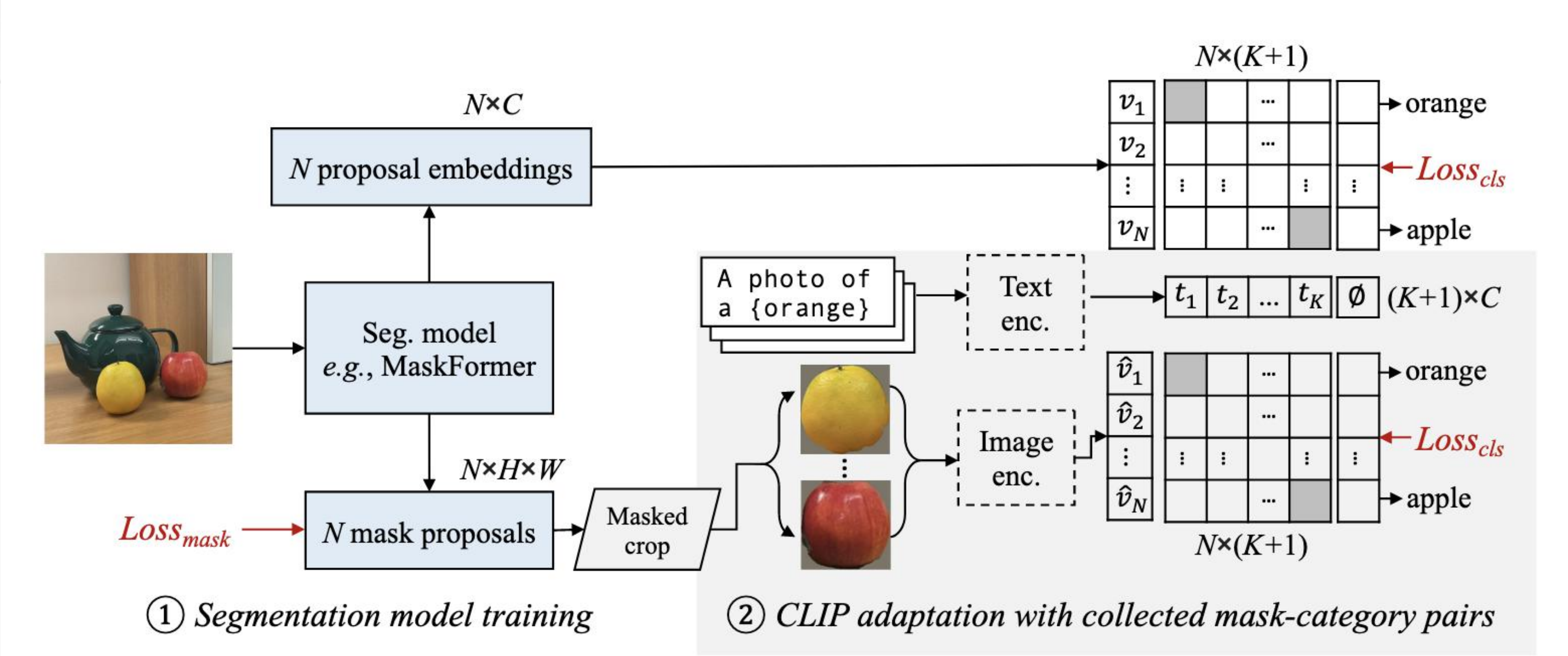
PART

Method

Overview


N×C
N proposal embeddings
Seg. model
e.g., MaskFormer
N×H×W
Loss_mask
N mask proposals
Masked crop
N×(K+1)
v_1
v_2
v_N
orange
Loss_cls
apple
A photo of a {orange}
Text enc.
t_1 t_2 ... t_K Ø
(K+1)×C
Image enc.
v̂_1
v̂_2
v̂_N
orange
Loss_cls
apple
N×(K+1)
① Segmentation model training
② CLIP adaptation with collected mask-category pairs

segmentation model



- mask result: N * H * W
- classification result： N * C

$$p_{i,k} = \exp(\sigma(v_i, t_k)/\tau) / \sum_k (\exp(\sigma(v_i, t_k)/\tau)).$$

segmentation model

	MaskFormer only	CLIP only	Ensemble
baseline	**19.6**	14.3	21.8
OVSeg (Ours)	**19.6**	**25.1**	**29.6**

- Training dataset : COCO STUFF (171 classes)

Adapt CLIP

- A straight method

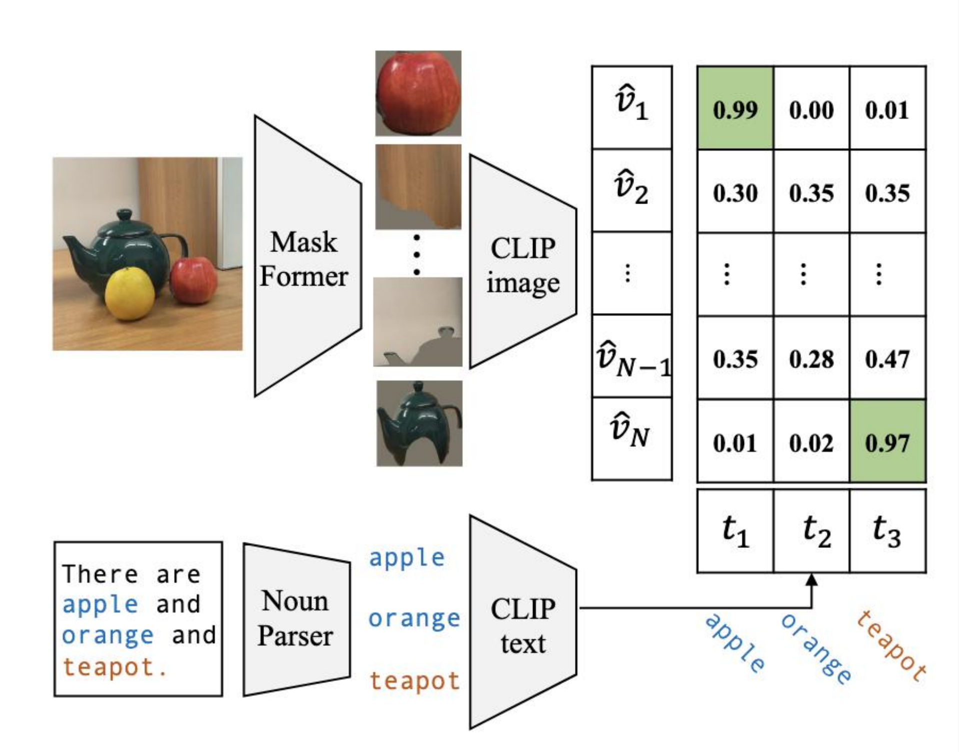
  --- fine tune CLIP on COCO-stuff

  --- CLIP overfits in the closed set

  --- Lack generalization ability

- Collecting diverse mask-category pairs from captions

Adapt CLIP

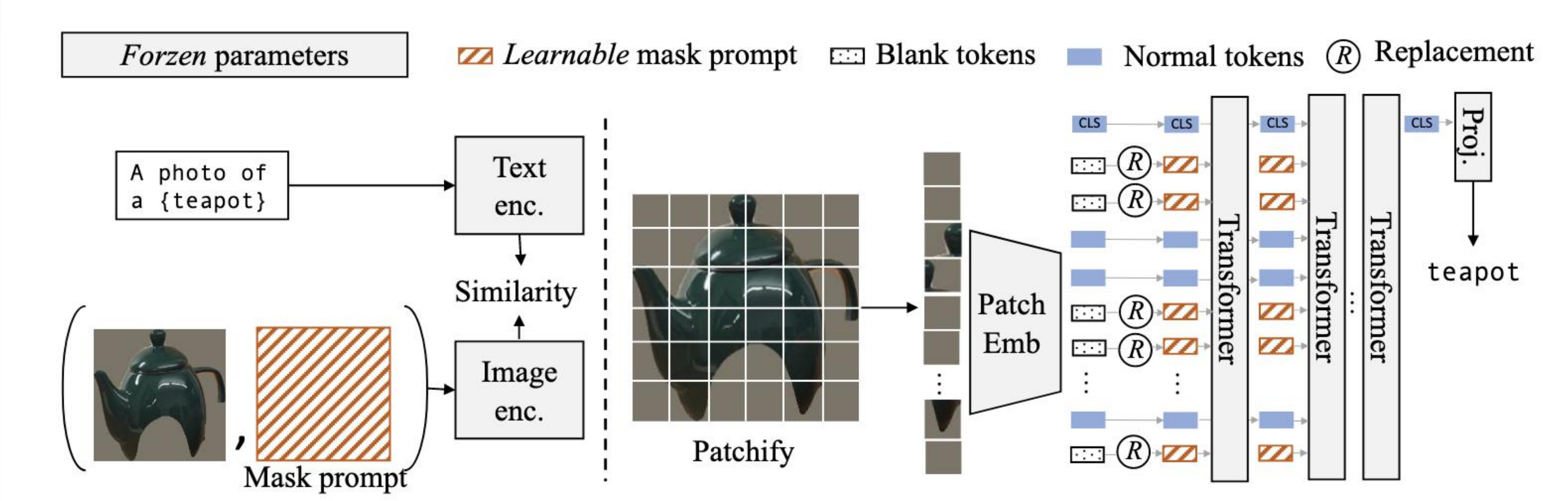


- COCO-Stuff -> COCO-Caption

- A self-labeling strategy

-- mask proposals
-- potential classes (get nouns)
-- matching (CLIP)

Mask Prompt Tuning

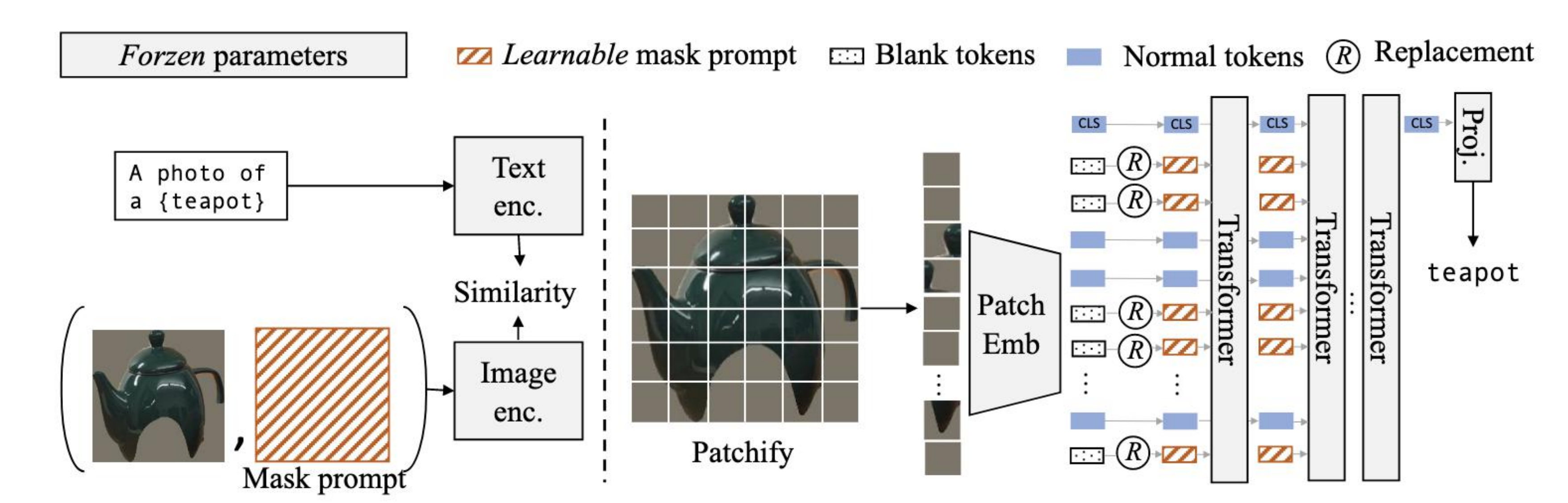


- Background is set to 0

-- token not contain useful information

-- domain shift

Mask Prompt Tuning



$$T \in \mathbf{R}^{N_p \times E}$$

$$M_p \in \{0,1\}^{N_p}$$

$$P \in \mathbf{R}^{N_p \times E}$$

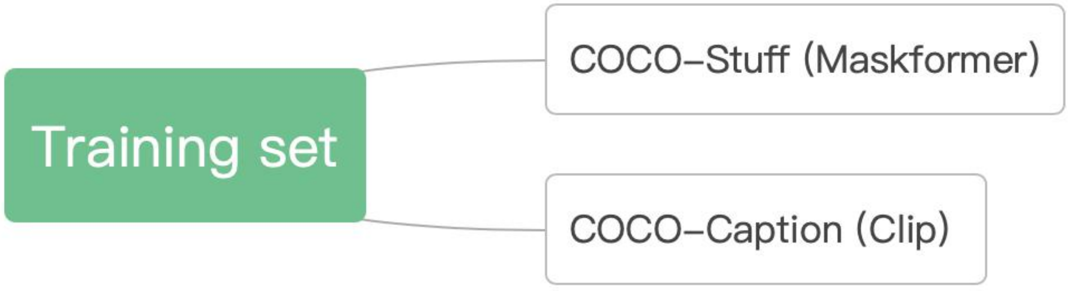
$$T \otimes M_p + P \otimes (1 - M_p)$$

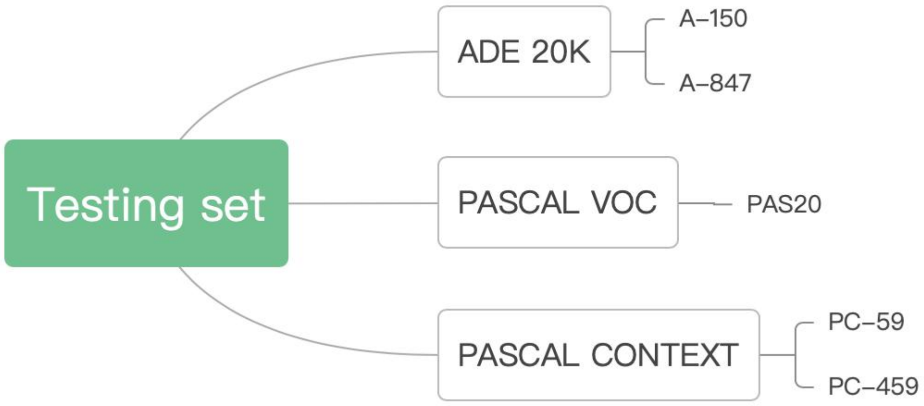
03

PART

Experiment

Dataset


Training set
COCO–Stuff (Maskformer)
COCO–Caption (Clip)


Testing set
ADE 20K
A–150
A–847
PASCAL VOC
PAS20
PASCAL CONTEXT
PC–59
PC–459

Results

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**
Supervised specialist models							
FCN [29]	FCN-8s	Same as test	-	-	29.4	37.8	-
Deeplab [6]	R-101	Same as test	-	-	-	45.7	77.7
SelfTrain [45]	Eff-L2	Same as test	-	-	-	-	90.0
MaskFormer [9]	R-101c	Same as test	17.4	-	46.0	-	-

Results

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**
Supervised specialist models							
FCN [29]	FCN-8s	Same as test	-	-	29.4	37.8	-
Deeplab [6]	R-101	Same as test	-	-	-	45.7	77.7
SelfTrain [45]	Eff-L2	Same as test	-	-	-	-	90.0
MaskFormer [9]	R-101c	Same as test	17.4	-	46.0	-	-

Results

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**
Supervised specialist models							
FCN [29]	FCN-8s	Same as test	-	-	29.4	37.8	-
Deeplab [6]	R-101	Same as test	-	-	-	45.7	77.7
SelfTrain [45]	Eff-L2	Same as test	-	-	-	-	90.0
MaskFormer [9]	R-101c	Same as test	17.4	-	46.0	-	-

Ablation studies

Case	Source		Statistics		A-847	A-150	PC-59
	Mask	Category	Pairs	Unique nouns			
Baseline	-	-	-	-	7.3	21.8	51.4
(1)	GT	GT	965K	171	5.3 (-2.0)	23.0 (+1.2)	**57.3** (+5.9)
(2)	GT	1 caption	440K	12K	7.9 (+0.6)	24.2 (+2.4)	53.2 (+1.8)
(3)	proposals	1 caption	440K	12K	**8.8** (+1.5)	**28.8** (+7.0)	55.7 (+4.3)
(4)	proposals	5 captions	1.3M	27K	**8.8** (+1.5)	28.6 (+6.8)	55.5 (+4.1)

- Ablation on mask-category pairs

Ablation studies

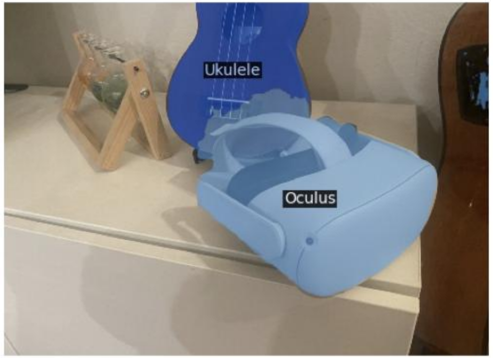
case	FT method MPT	FT method full	A-847	A-150	PC-59
Baseline			7.3	21.8	51.4
(a)	✓		8.4 (+1.1)	26.5 (+4.7)	55.4 (+4.0)
(b)		✓	8.8 (+1.5)	28.8 (+7.0)	**55.7** (+4.3)
(c)	✓	✓	**9.0** (+1.7)	**29.6** (+7.8)	**55.7** (+4.3)

combination	A-847	A-150
FT ->MPT (default)	**9.0**	**29.6**
MPT ->FT	8.5 (-0.5)	28.1 (-1.5)
FT + MPT sim.	8.8 (-0.2)	29.0 (-0.6)

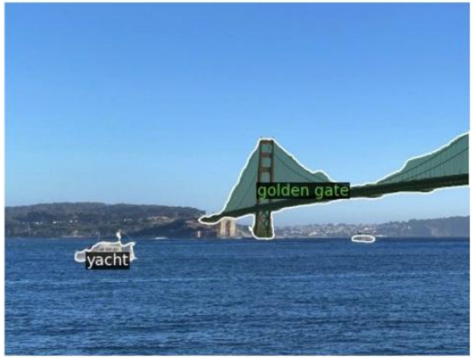
- Ablation on mask prompt tuning

Visualization

Query: `saturn V, blossom`



Query: `Oculus, Ukulele`



Query: `golden gate, yacht`

Visualization

GT: building Pred: skycraper

GT: rail Pred: road

04

PART

Conlusion

conclusion

- 将Clip用到开放域任务中如何保持其泛化性？（利用caption构造）
- Finetune Clip时候使用prompt方法不破坏其参数

THANKS

感谢聆听！