FUDAN UNIVERSITY

JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

Learning Open-Vocabulary Semantic Segmentation Models from Natural Language Supervision

Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, Weidi Xie

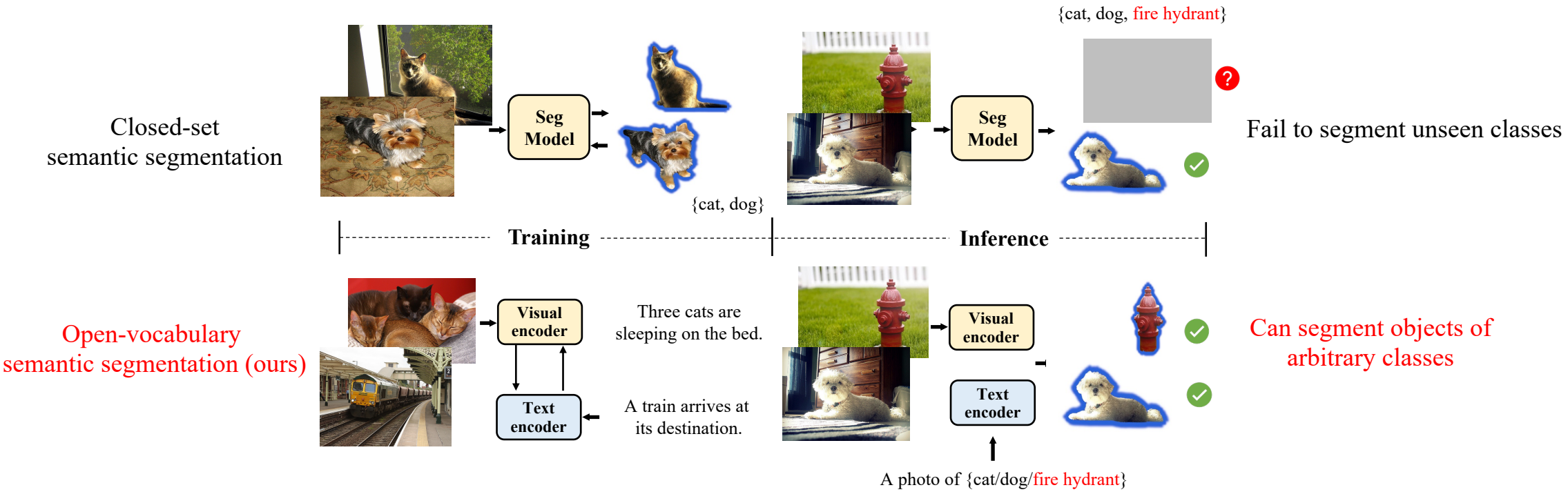
Fudan University, Shanghai AI Lab, Shanghai Jiaotong University

Paper Tag: TUE-AM-279

Jun 19th , 2023

Objective

- Learning an open-vocabulary semantic segmentation model

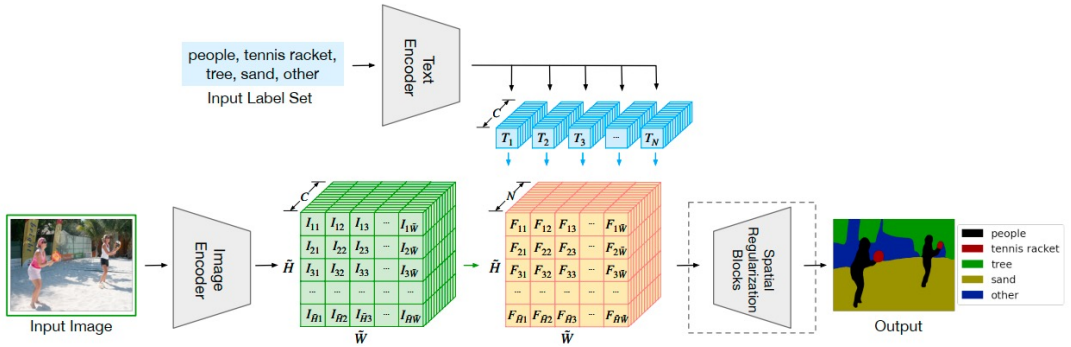


✓ With natural language supervision only, i.e., without any mask annotations

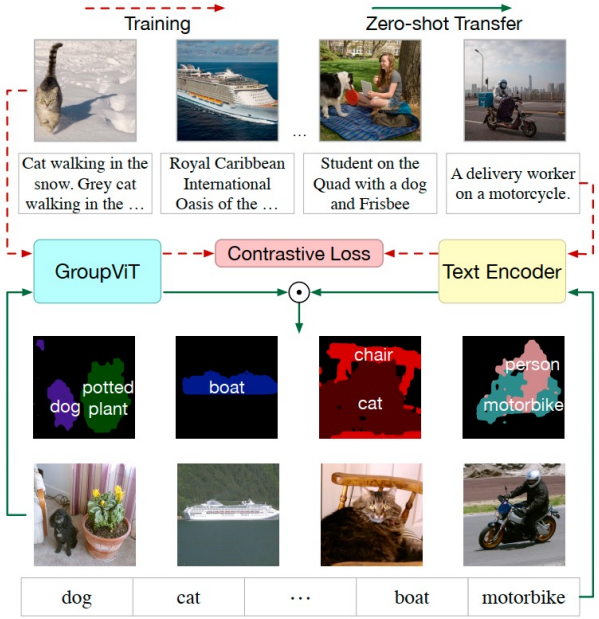
✓ Zero-shot transfer to any segmentation dataset

Related Works

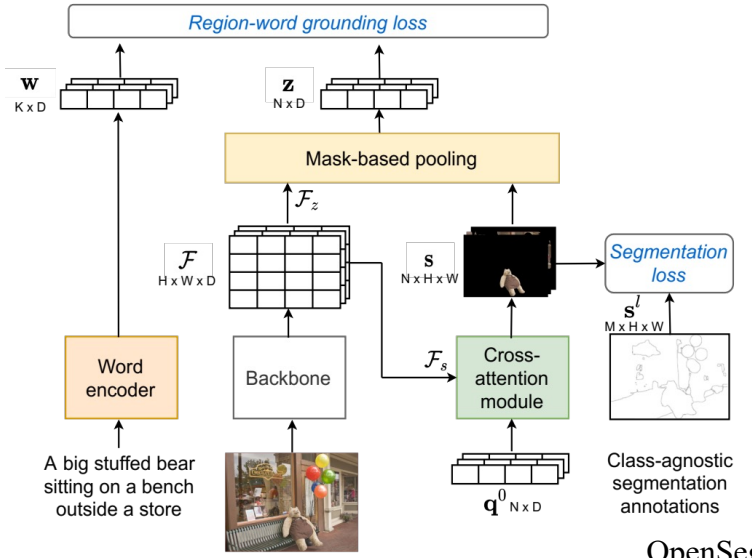
- Zero-shot/Language-guided semantic segmentation



LSeg. ICLR 22



GroupViT. CVPR 22

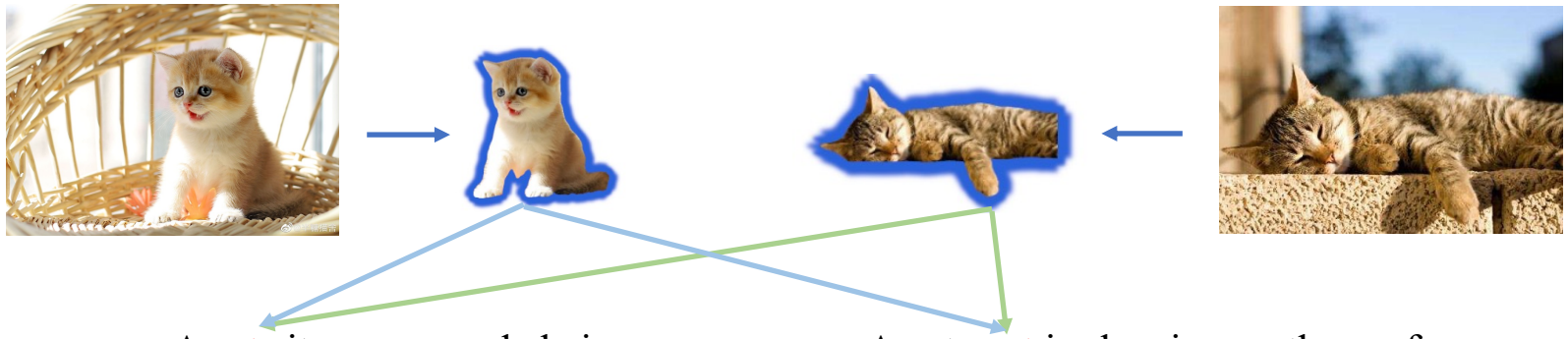


OpenSeg. ECCV 22

	Open-vocab ability?	Mask-free training?	Zero-shot transfer?	Data-efficient training?
LSeg	✅	❌	❌	✅
OpenSeg	✅	❌	❌	✅
GroupViT	✅	✅	✅	❌
Ours	✅	✅	✅	✅

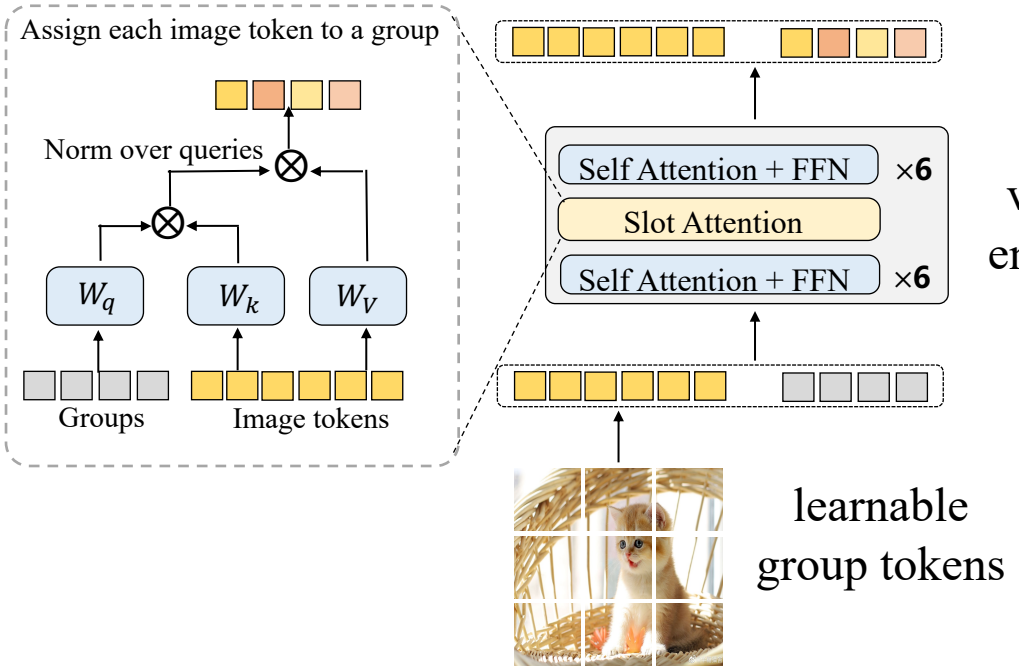
Challenges & Intuitions

- Challenges
  - The captions (e.g. CC12M/LAION) only provide coarse, image-level descriptions
  - Large diversity of web-collected data
- Our intuition
  - Exploiting the visual invariance between different images

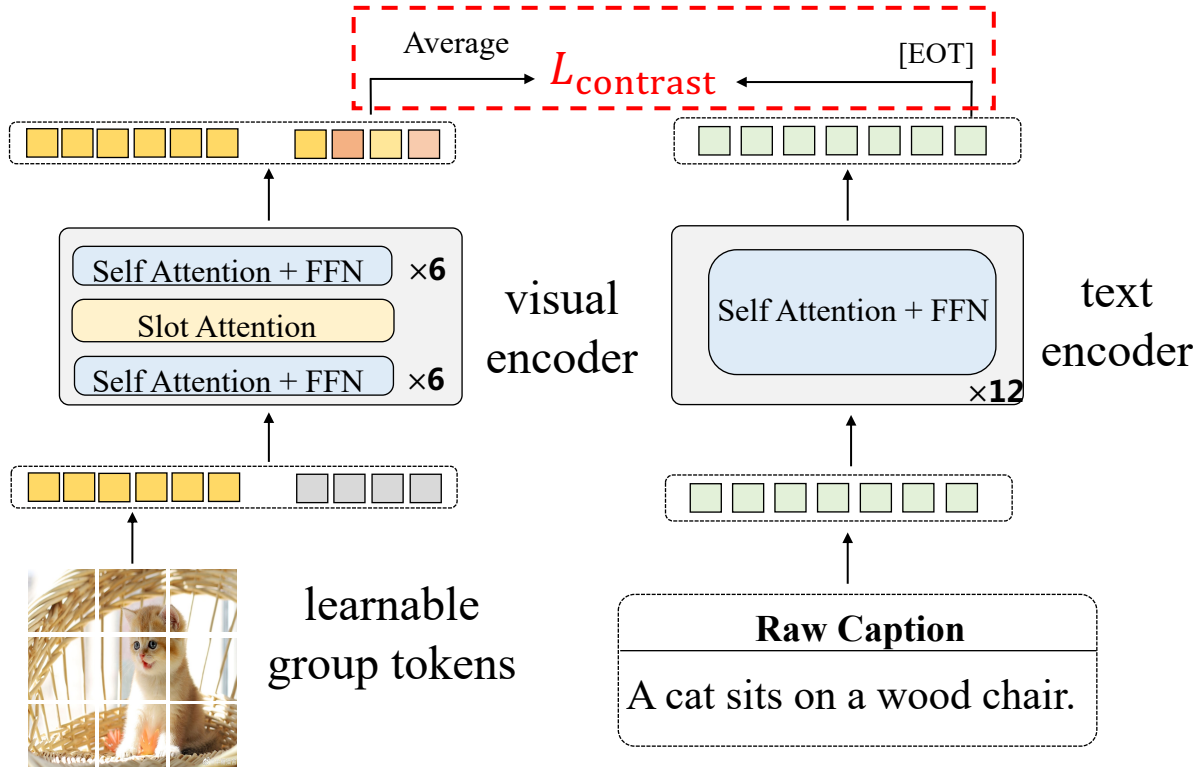


- Overview

  - Step1: Visual grouping (patch to groups)



  - Step2: Group-text alignment (groups to caption)

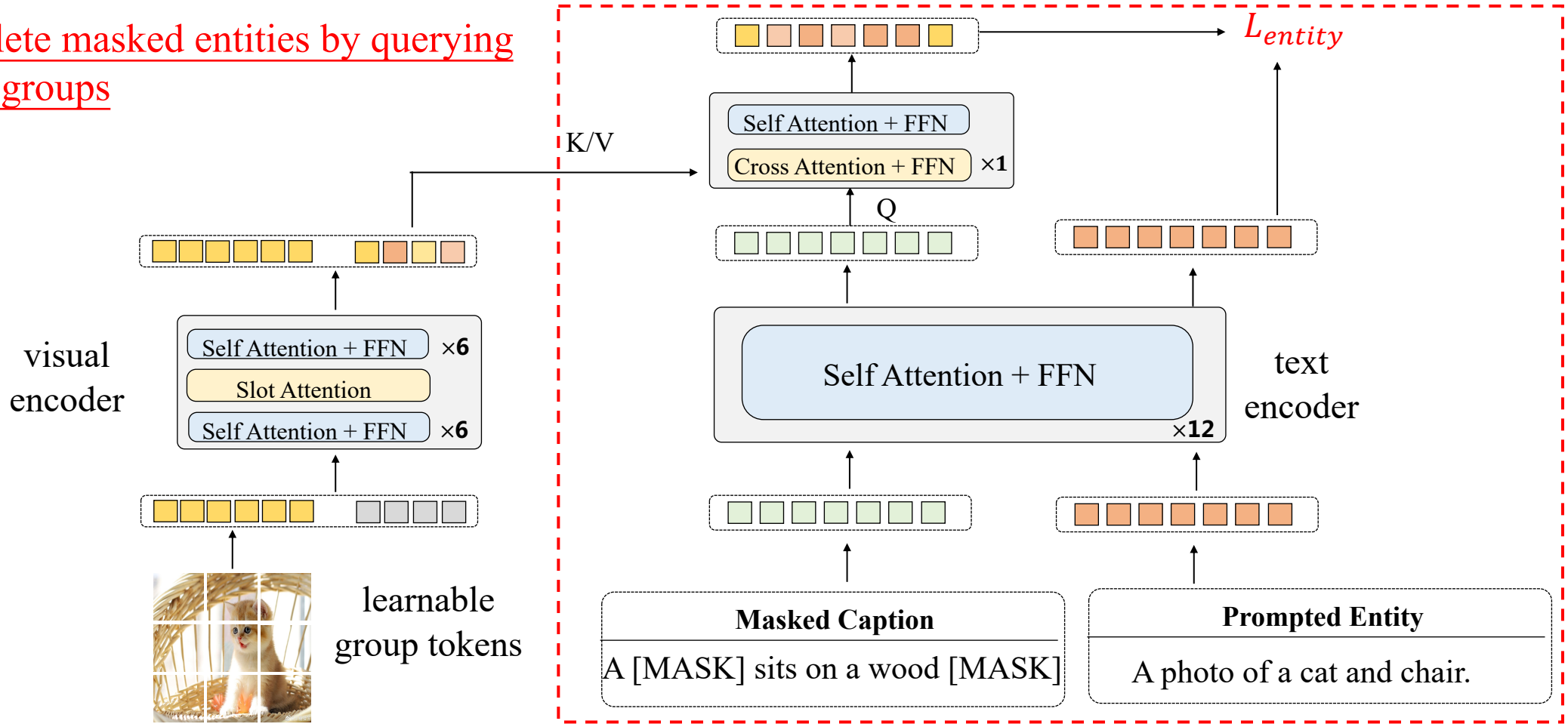


[1] Xu et al. GroupViT: Semantic Segmentation Emerges from Text Supervision, In CVPR 2022

- Proxy tasks to learn visual invariance

Task1: Masked entity completion

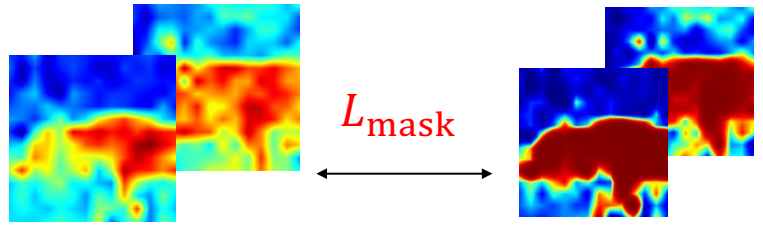
➢ Complete masked entities by querying visual groups



- Proxy tasks to learn visual invariance

Task2: Cross-image mask consistency

➢ Consistent mask predictions between images that contain shared entities (e.g., cat)

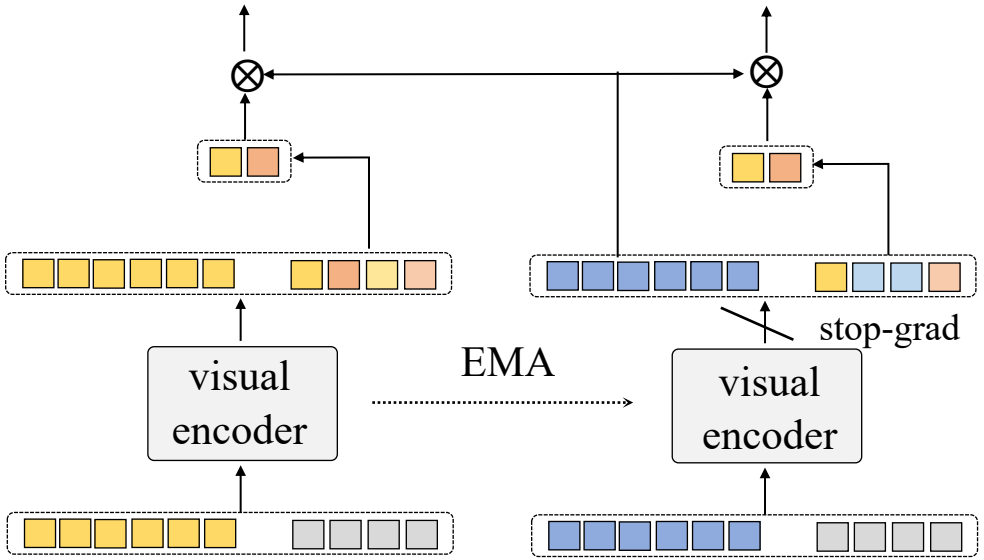


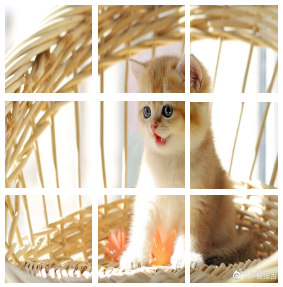
Pipeline

1) Find entity-specific sub-groups

2) Solve bipartite matching between two sub-groups

3) Pairwise Dice loss between two sets of masks





Experiment settings

➢ Pre-training dataset

- **CC4M**: A subset of CC12M that contain 100 commonly appeared entities (e.g. person, shirt, cat, dog, bus, cup, knife, plant, …), and non visual entities (e.g. view, illustration, night) are discarded.

  Researchers can leverage useful visual entities in CC4M under limited computation resources.

➢ Zero-shot evaluation

- PASCAL VOC, PASCAL Context, COCO Stuff, ADE20K, …

➢ Network architecture

- Visual encoder: ViT-B / ViT-S
- Text encoder: BERT-base

Results

- On PASCAL VOC, our model (w/ zero-shot transfer) outperforms the model using supervised finetuning.
- Our model outperforms the SOTA by using only 3% pre-trained data (4M vs. 134M)

Method	Backbone	Pretrain dataset	Supervision	Zero-shot transfer	Downstream datasets			
					PASCAL VOC	PASCAL Context	COCO Object	ADE20K
DeiT [46]	ViT-S	IN-1K	class label	✗	53.0	35.9	-	-
MoCo [10]	ViT-S	IN-1K	self	✗	34.3	21.3	-	-
DINO [6]	ViT-S	IN-1K	self	✗	39.1	20.4	-	-
MoCo [10]	ViT-S	CC12M+YFCC15M	self	✗	36.1	23.0	-	-
DINO [6]	ViT-S	CC12M+YFCC15M	self	✗	37.6	22.8	-	-
ViL-Seg [32]	ViT-B	CC12M	self+text	✓	33.6	15.9	-	-
GroupViT [53]	ViT-S	CC12M	text	✓	41.1	-	-	-
GroupViT [53]	ViT-S	CC12M+YFCC15M	text	✓	51.2	22.3	20.9	-
ViewCo [43]	ViT-S	CC12M+YFCC15M	self+text	✓	52.4	**23.0**	23.5	-
CLIPpy [42]	ViT-B	HQITP-134M	text	✓	52.2	-	25.5†	13.5
GroupViT* [53]	ViT-S	CC4M	text	✓	19.8	8.8	9.1	3.4
GroupViT* [53]	ViT-B	CC4M	text	✓	25.8	11.3	10.7	3.6
GroupViT* [53]	ViT-S	CC12M	text	✓	40.2	18.7	17.7	**6.2**
OVSegmentor (ours)	ViT-S	CC4M	self+text	✓	44.5	18.3	19.0	4.3
OVSegmentor (ours)	ViT-B	CC4M	self+text	✓	**53.8**	20.4	**25.1**	5.6

Results

- Comparison with zero-shot semantic segmentation methods

Method	Pretrain dataset	Seen	Unseen	VOC	Context
SPNet [51]	-	✔	✗	15.6	4.0
ZS3 [5]	-	✔	✔	17.7	7.7
GaGNet [22]	-	✔	✔	29.9	15.0
SIGN [14]	-	✔	✔	28.9	14.9
Joint [2]	-	✔	✗	32.5	-
ZegFormer [17]	CLIP400M	✔	✗	63.6	
MaskCLIP+ [60]	CLIP400M	✔	✗	86.1	66.7†
ViL-Seg [32]	CC12M	✗	✗	37.3	18.9
GroupViT [53]	CC12M+YFCC15M	✗	✗	43.7	51.3
GroupViT* [53]	CC4M	✗	✗	22.4	24.5
GroupViT* [53]	CC12M	✗	✗	33.1	45.3
OVSegmentor	CC4M	✗	✗	**46.6**	**54.5**

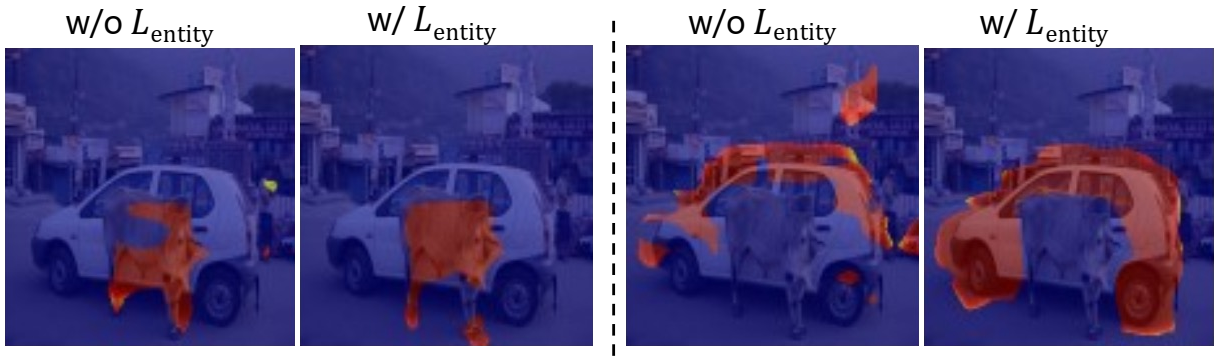
Table 2. **Comparison with zero-shot segmentation methods on unseen classes.** Seen/Unseen denotes whether the model is trained on seen/unseen classes. Our method outperforms most zero-shot segmentation models even without training on seen classes. † indicates a different set of unseen classes.

- Our model outperforms most zero-shot segmentation models even without training with mask annotations on the seen classes.

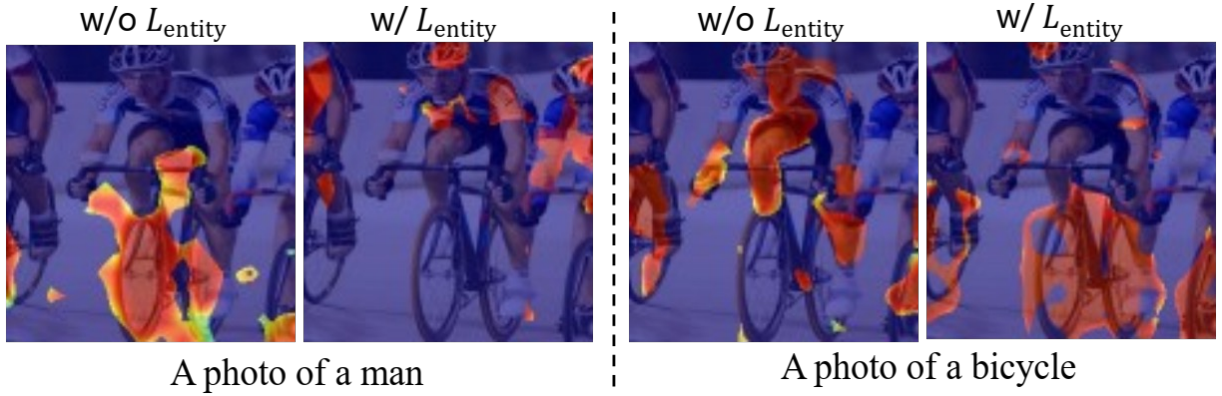
Results

- Effect of the masked entity completion loss

  - Improvement on visual grouping



  - Improvement on group-text matching



  - Comparison with variants

Masking objective	PASCAL VOC	PASCAL Context
All entities (ours)	**48.9**	**19.9**
Single entity	47.0	18.5
All nouns	45.4	17.4
Multi-label contrastive	44.5	17.0
MLM (w/ groups)	42.8	16.3
MLM (w/o groups)	36.3	14.6

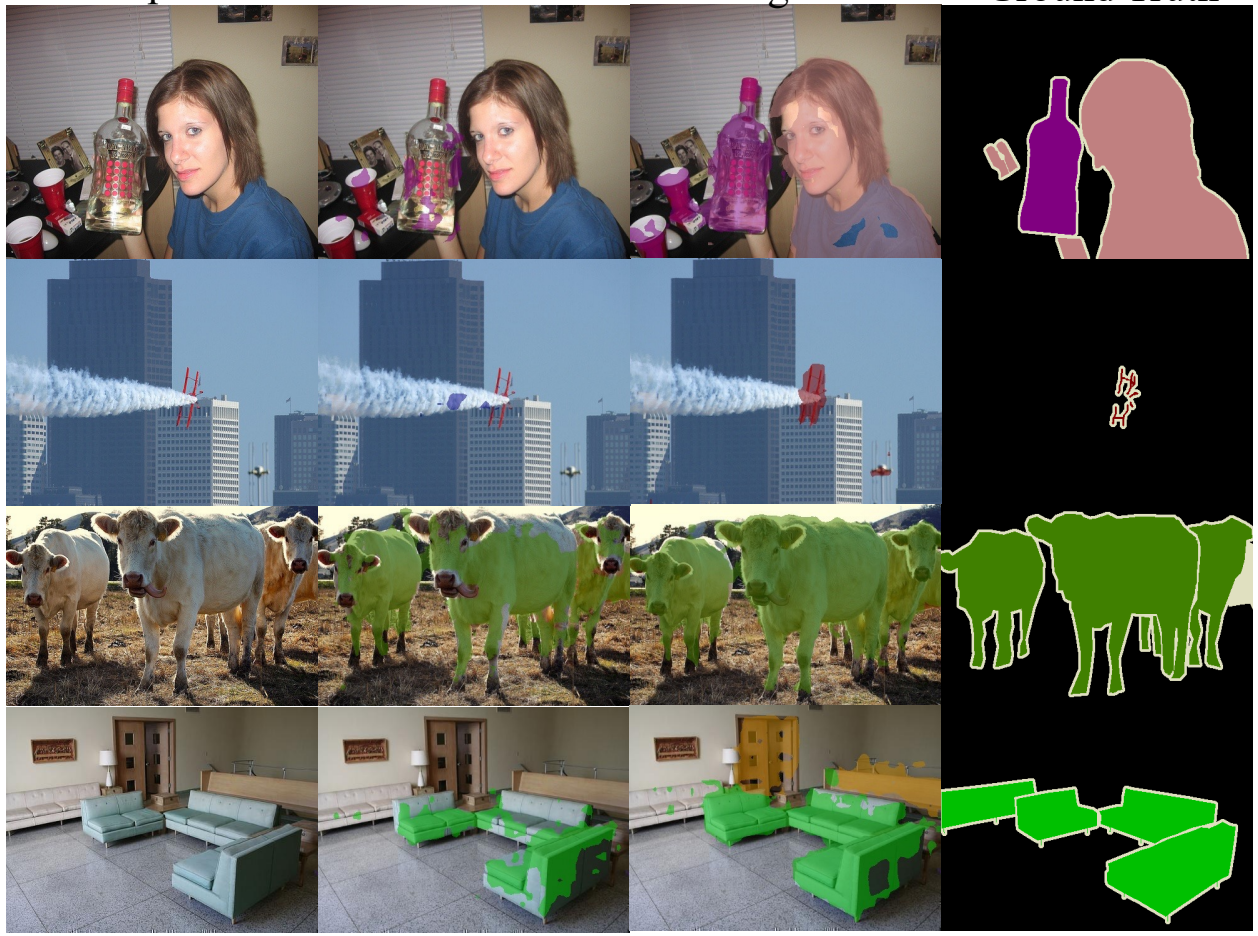
Example:
The cat is sleeping under the tree on a sunny day.

1. Ours (all entities) : The [MASK] is sleeping under the [MASK] on a sunny day.
2. Single entity: The [MASK] is sleeping under the tree on a sunny day.
3. All nouns: The [MASK] is sleeping under the [MASK] on a sunny [MASK].

Visualisation

On PASCAL VOC

Input | Baseline | OVSegmentor | Ground Truth



On COCO Object

Input | Baseline | OVSegmentor | Ground Truth

Conclusion

Follow ups

- Handling background classes in the open-vocabulary segmentation.
- Federated training on hybrid datasets (segmentation + caption)
- SAM in the loop

Paper: https://arxiv.org/abs/2301.09121

Project page: https://jazzcharles.github.io/OVSegmentor/

Code: https://github.com/Jazzcharles/OVSegmentor

Email: jilanxu21@m.fudan.edu.cn

OVSegmentor

THANKS FOR LISTENING, PLEASE COMMENT

Jilan Xu