Class Distribution-induced Attention Map for Open-vocabulary Semantic Segmentations



Dong Un Kang[1] **Hayeon Kim**[1] **Se Young Chun**[1,2†]

Intelligent Computational imaging Lab (ICL)

[1]Dept. of Electrical and Computer Engineering, [2]INMC & IPAI

Seoul National University, Republic of Korea

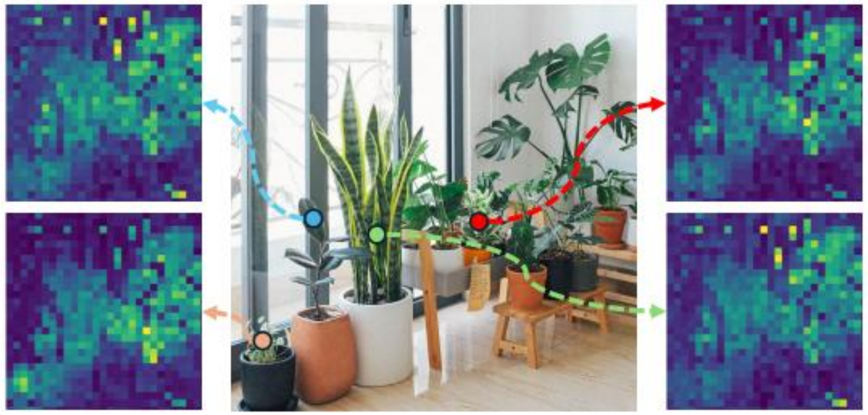
† Corresponding author

Our poster session on **Thursday, April 24, from 3:00 p.m. to 5:30 p.m.**

Background: Recent works on open-vocabulary semantic segmentation

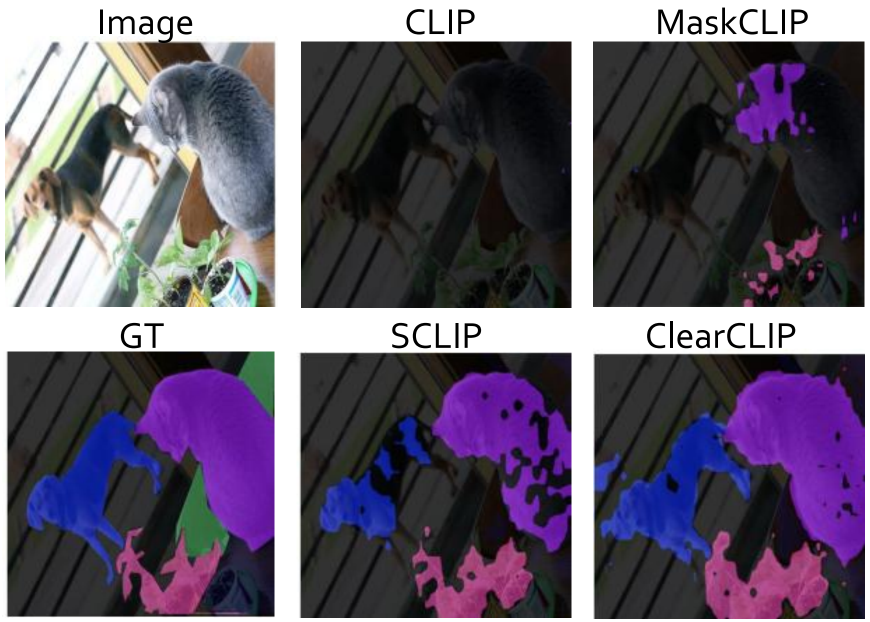
✓ By refining CLIP's noisy attention map, recent CLIP-based methods have enabled open-vocabulary semantic segmentation without additional training.

✓ However, they still struggle to accurately localize the target object and often produces noisy predictions.



Final layer attention maps in image encoder of CLIP

* SCLIP: Wang, Feng, et al. ECCV, 2024

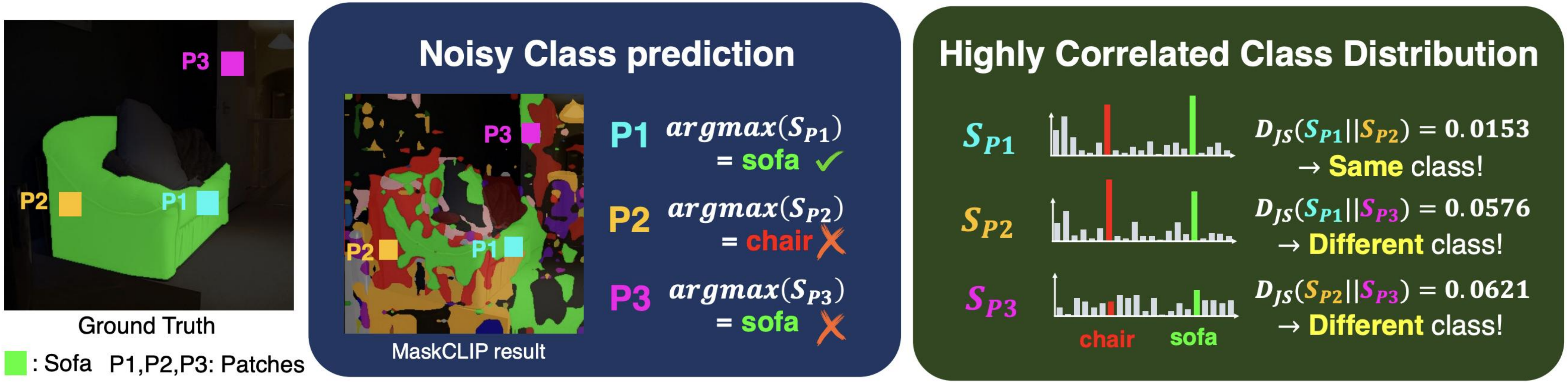


Semantic segmentation results of recent works

*ClearCLIP: Lan, Mengcheng, et al. ECCV, 2024

Motivation of our work

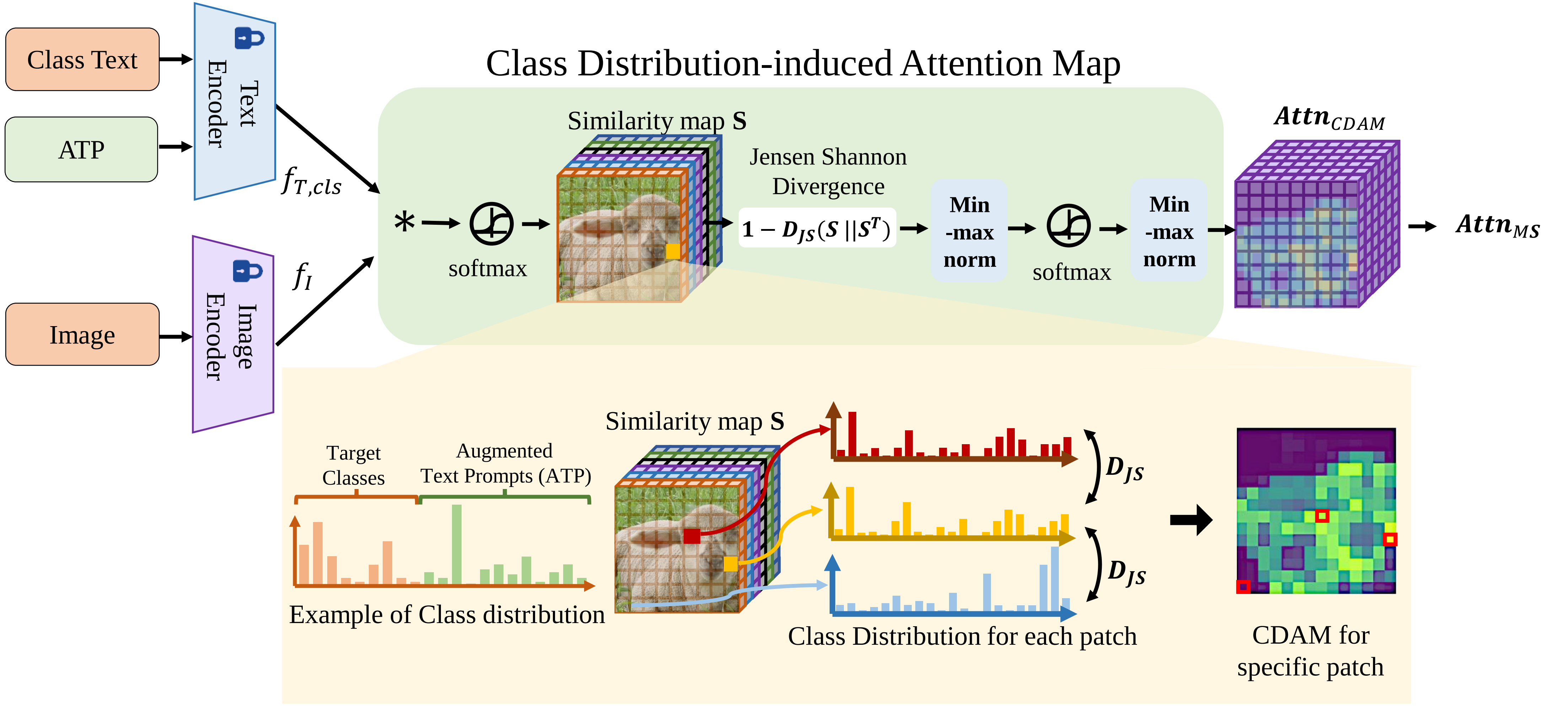
✓ Recent CLIP-based methods often make noisy predictions for each patch in an image. Nevertheless, we observe that patches belonging to the same object tend to have highly similar class distributions.



The similarity of the class distribution patches between **P1** and **P2** (same class) is more similar than between **P1** and **P3** (different class).

Method: Class Distribution-induced Attention Map (CDAM)

1) Generate the **similarity map S** which is the initial prediction of existing methods.
2) Calculate the **Jensen Shannon divergence** between class distribution for constructing CDAM.



Experimental results: Quantitative comparison

Comparison of open-vocabulary semantic segmentation with SOTA models

✓ Consistently improves upon prior works, achieving up to a 22.8% increase in mIoU.

*Evaluation metric : mIoU

Method	Pre-trained Model	Extra Training	VOC21	Context60	COCO-Obj	Avg.
CLIP-based training-free methods						
CLIPSurgery (Li et al., 2023)	CLIP	✗	-	29.3	-	-
CLIP-DIY (Wysoczańska et al., 2024)	CLIP+DINO	✗	59.0	-	30.4	-
CaR† (Sun et al., 2024)	CLIP	✗	**59.4**	25.0	33.2	39.2
MaskCLIP† (Zhou et al., 2022)	CLIP	✗	33.1	23.3	24.8	27.1
MaskCLIP+**CDAM**	CLIP	✗	55.9 (+22.8)	30.5 (+7.2)	34.3 (+9.5)	40.2 (+13.1)
SCLIP† (Wang et al., 2023)	CLIP	✗	50.5	25.8	31.3	35.9
SCLIP+**CDAM**	CLIP	✗	59.0 (+8.5)	30.4 (+4.5)	34.5 (+3.0)	41.3 (+5.4)
ClearCLIP† (Lan et al., 2024)	CLIP	✗	50.7	27.8	33.0	37.2
ClearCLIP+**CDAM**	CLIP	✗	57.6 (+6.9)	29.8 (+2.0)	34.5 (+1.5)	40.6 (+3.4)
GEM† (Bousselham et al., 2024)	CLIP	✗	52.1	28.1	33.8	38.0
GEM+**CDAM**	CLIP	✗	58.7 (+6.6)	**30.6** (+2.5)	**35.2** (+1.4)	**41.5** (+3.5)

Benchmark datasets *with* background class

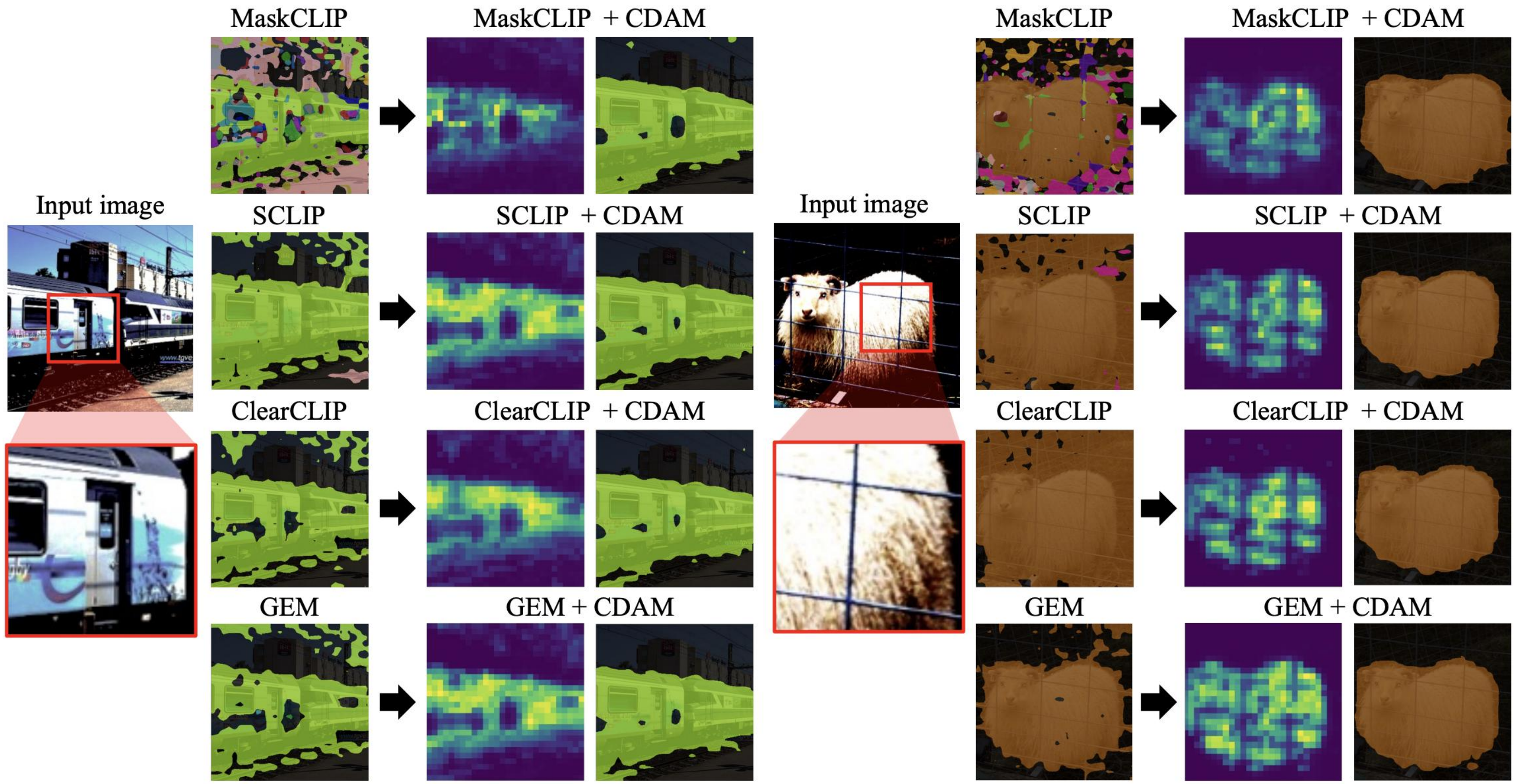
Method	Pre-trained Model	Extra Training	COCO-Stf	CityScapes	ADE20K	Avg.
CLIP-based training-free methods						
MaskCLIP† (Zhou et al., 2022)	CLIP	✗	16.5	23.8	12.2	17.5
MaskCLIP+**CDAM**	CLIP	✗	24.5 (+8.0)	**27.6** (+3.8)	**17.8** (+5.6)	**23.3** (+5.8)
SCLIP† (Wang et al., 2023)	CLIP	✗	21.1	19.7	14.6	18.5
SCLIP+**CDAM**	CLIP	✗	24.5 (+3.4)	24.6 (+4.9)	17.2 (+2.6)	22.1 (+3.6)
ClearCLIP† (Lan et al., 2024)	CLIP	✗	23.9	20.8	16.6	20.4
ClearCLIP+**CDAM**	CLIP	✗	24.6 (+0.7)	21.7 (+0.9)	17.1 (+0.5)	21.1 (+0.7)
GEM† (Bousselham et al., 2024)	CLIP	✗	23.7	21.2	15.7	20.2
GEM+**CDAM**	CLIP	✗	**24.8** (+1.1)	23.7(+1.5)	17.2 (+1.5)	21.9 (+1.7)

Benchmark datasets *without* background class

Experimental results: Qualitative results

Segmentation results achieved by integrating CDAM with prior works

✓ By generating high-quality attention map (CDAM) from initial predictions of prior works, we synergistically enhance localization accuracy and reduces noise prediction through the application of CDAM into the final layer of CLIP.



Intelligent Comput. imaging Lab.

ICLR 2025

Thank you!

Class Distribution-induced Attention Map for Open-vocabulary Semantic Segmentations

Please visit our poster session on **Thursday, April 24, from 3:00 p.m. to 5:30 p.m.**

Paper: https://openreview.net/pdf?id=CMqOfvD3tO
Project page: https://janeyeon.github.io/cdamclip/

Project page

Paper

Acknowledgement: This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF2022M3C1A309202211), AI-Bio Research Grant through Seoul National University and Samsung Electronics Co., Ltd (IO221004-02674-01). Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.