

ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference

Mengcheng Lan¹, Chaofeng Chen¹, Yiping Ke², Xinjiang Wang³,
Litong Feng^{3*}, and Wayne Zhang³

¹ S-Lab, Nanyang Technological University

² CCDS, Nanyang Technological University ³ SenseTime Research
lanm0002@e.ntu.edu.sg {chaofeng.chen, ypke}@ntu.edu.sg
{wangxinjiang, fenglitong, wayne.zhang}@sensetime.com
<https://github.com/mc-lan/ClearCLIP>

- Problem / objective
 - CLIP 사용해서 Open-Vocabulary Semantic Segmentation
- Contribution / Key idea
 - Vision encoder의 마지막 layer에 3가지 수정을 함
 - 1. Residual connection 제거
 - 2. Self-self attention 적용
 - 3. Feed-forward network 제거

- **Semantic segmentation using CLIP**

- ❑ Noisy 하다.
- ❑ 이 noise는 어디서 왔으며, 어떻게 발생하였는가?

$$\mathcal{M} = \arg \max_c \cos(X_{\text{dense}}^{\text{visual}}, X^{\text{text}}). \quad (4)$$



Image



CLIP

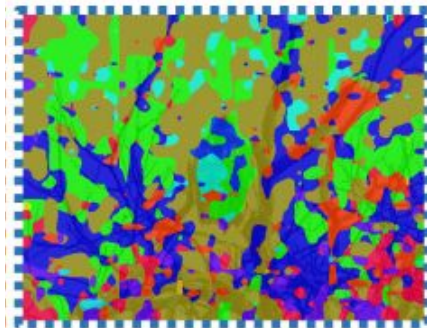
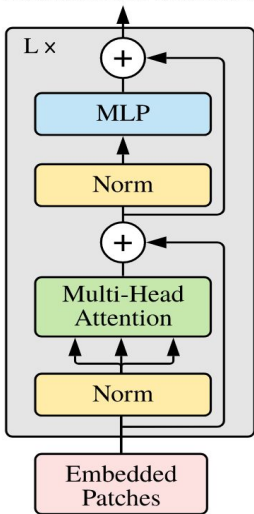
- **Analysis of segmentation results using CLIP**

[질문1] 이 noise는 어디서 왔는가?

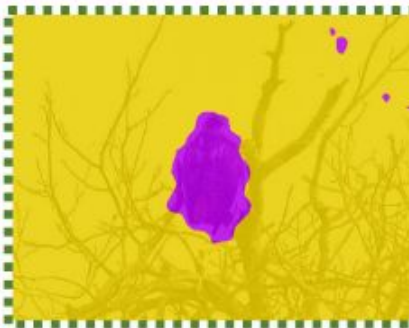
[정답] Residual connection

[근거] CLIP vision encoder에서 최종 layer의 (1) residual connection (2) attention output 비교 결과

Transformer Encoder



Residual connection



Attention output



CLIP

● Analysis of segmentation results using CLIP

[질문2] 그렇다면, residual connection에서 noise가 어떻게 발생하였는가?

[정답] CLIP의 high-level supervision이 residual latent space에서 global feature direction을 강조하여, residual features에서의 noise가 local feature vector들이 구분하기 어렵게 만들.

[근거] (1) Residual connection과 attention output 간에 엔트로피 차이 존재.

- Residual connection의 엔트로피 $\rightarrow 0$, Attention output의 엔트로피 $\rightarrow 1$

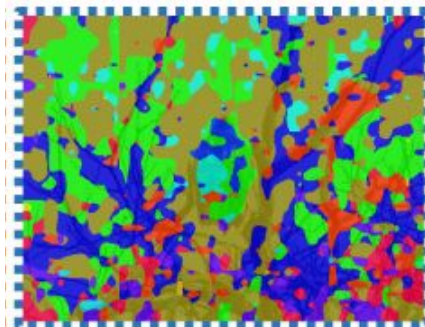
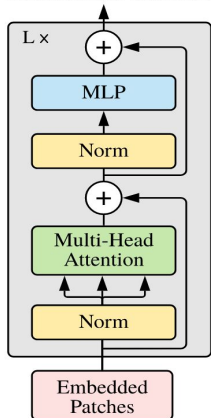
- I.E., Residual connection map의 값 분포가 특정 채널/위치에 몰려있다.

- I.E., CLIP의 최종 output도 residual connection과 비슷한 속성으로 됨.

(2) Residual connection map의 대부분 feature vector들의 peak값들이 특정 채널에 몰려있음.

- I.E., Latent space에서 이 feature vector들의 방향이 유사하여, cosine similarity로 각 spatial feature vector들을 구분하기가 어려워, noise 발생

Transformer Encoder



Residual connection



Attention output



CLIP

● Noise가 어디서 왔는가?

□ 실험 결과

(1) 'X_attn의 크기': mIoU와 양의 상관관계.

(2) 'X_res의 크기': CLIP-B/16에서 CLIP-L/14에서보다 훨씬 작은 값을 가진다.

(3) Attention 수정: mIoU가 q-k baseline보다 CLIP-B/16에서는 일관되게 높지만, CLIP-L/14에서는 그렇지 않다.

□ 가설

Attention 수정은 X_res의 영향(norm값)이 적을 때 효과적이다.

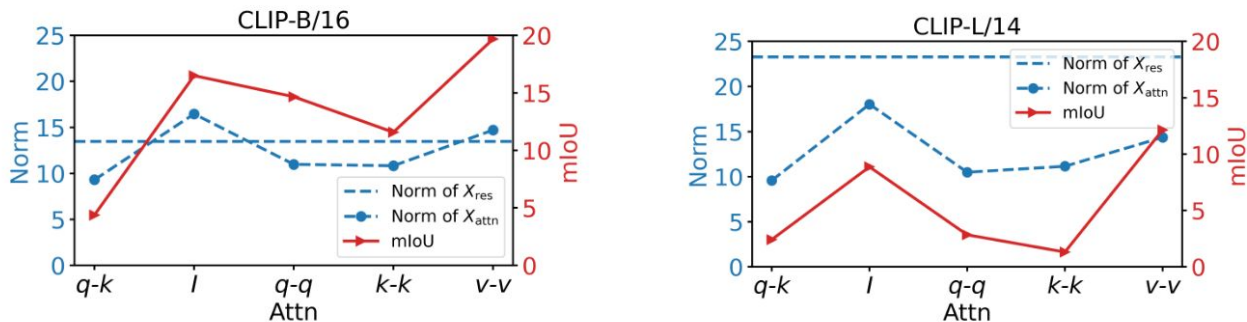


Fig. 2: Comparison of norms and mIoU of different attention mechanisms for CLIP-B/16 (left) and CLIP-L/14 (right). The norm curve of X_{attn} shows a positive correlation with the mIoU curve. A larger norm of X_{res} in CLIP-L/14 impedes the enhancement of performance through the revision of attention mechanisms.

- Noise가 어디서 왔는가?

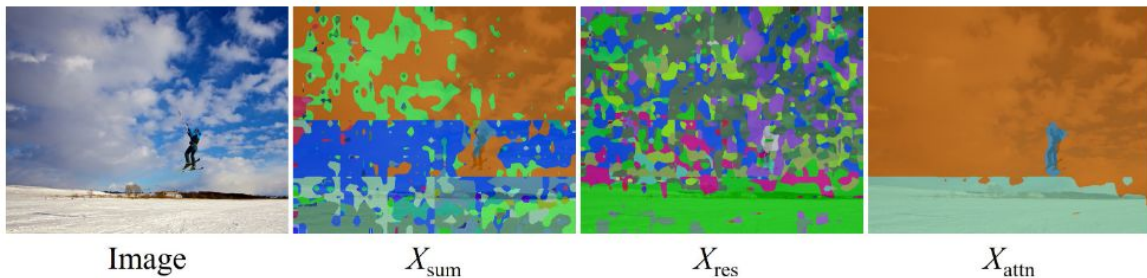
- ❑ 가설 검증

- (1) X_{res} 의 mIoU $\rightarrow 0$,

- (2) X_{attn} 의 mIoU $> X_{sum}$ 의 mIoU 인 것을 보아,
residual connection은 image segmentation에 도움이 전혀 안되고 있다.

- ❑ 결론

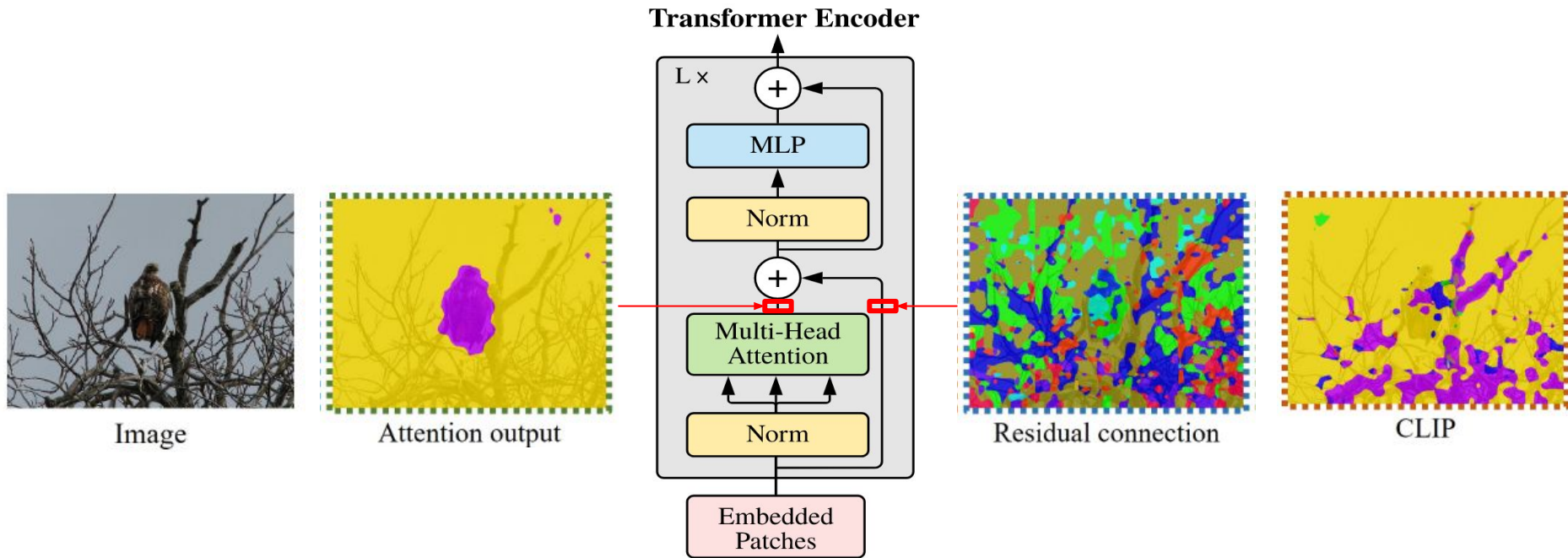
- CLIP을 사용했을때 segmentation map의 noise는 residual connection에서 비롯되었다.



| Features | mIoU |
|------------|------|
| X_{sum} | 4.4 |
| X_{res} | 0.01 |
| X_{attn} | 11.6 |

Fig. 3: Open-vocabulary semantic segmentation using different feature maps of CLIP-B/16 model on the COCOStuff dataset. A visualization of an example (left) and quantitative results (right).

- Noise가 어디서 왔는가?



- Noise가 어떻게 발생하였는가 ?

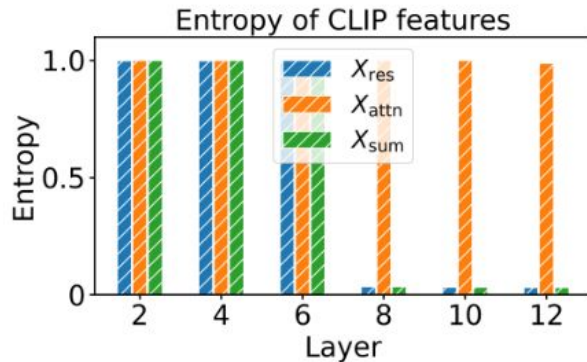
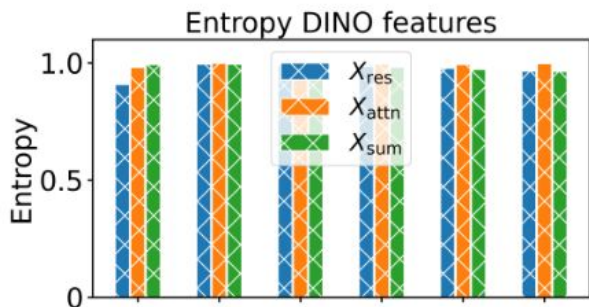
- 실험결과

(1) DINO-B/16의 feature: 레이어 상관없이 엔트로피 값 일정하게 유지,

CLIP-B/16의 feature: 레이어가 깊어짐에 따라 X_{res} 와 X_{sum} 의 엔트로피 값 급격히 (거의 0 수준까지) 감소.

-> I.E., CLIP-B/16의 깊은 레이어의 X_{res} 와 X_{sum} 에 peak값들이 존재한다는 것.

$$H(X^L) = -\frac{1}{\log(hw \times d)} \sum_{i,j} p(X_{i,j}^L) \log p(X_{i,j}^L), \quad p(X_{i,j}^L) = \frac{e^{X_{i,j}^L}}{\sum_{m,n} e^{X_{m,n}^L}}, \quad (5)$$

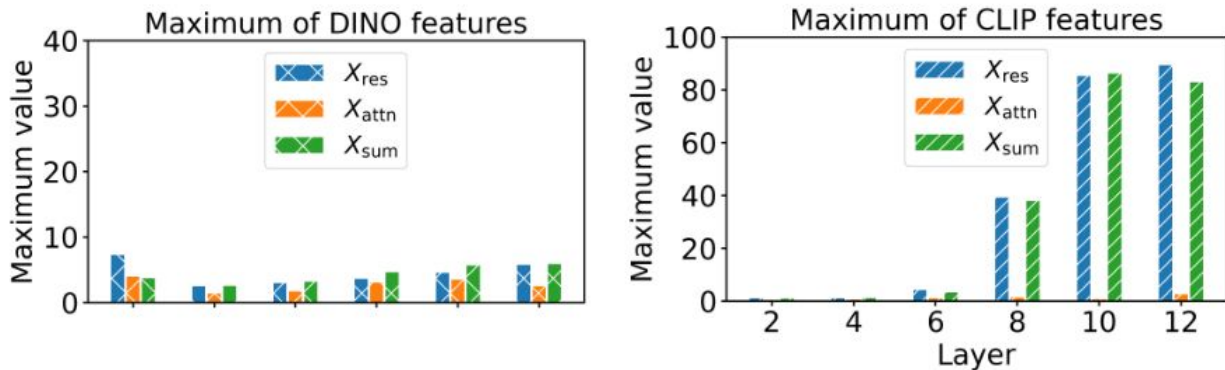


(a) Entropy

- Noise가 어떻게 발생하였는가 ?

- 실험결과

- (2) DINO-B/16의 feature: 레이어 변환에 따라 최대값이 (10 이하로) 안정적으로 유지,
CLIP-B/16의 feature: 레이어가 깊어짐에 따라 X_{res} 와 X_{sum} 의 최대값이 급격히 (초기보다 90배) 증가.
-> 이것이 CLIP-B/16의 레이어가 깊어짐에 따라 X_{res} 와 X_{sum} 의 엔트로피 값 급격히 감소한 이유.
-> I.E., X_{res} 의 값 분포가 특정 채널/위치에 몰려있다.



(b) Maximum

- Noise가 어떻게 발생하였는가 ?

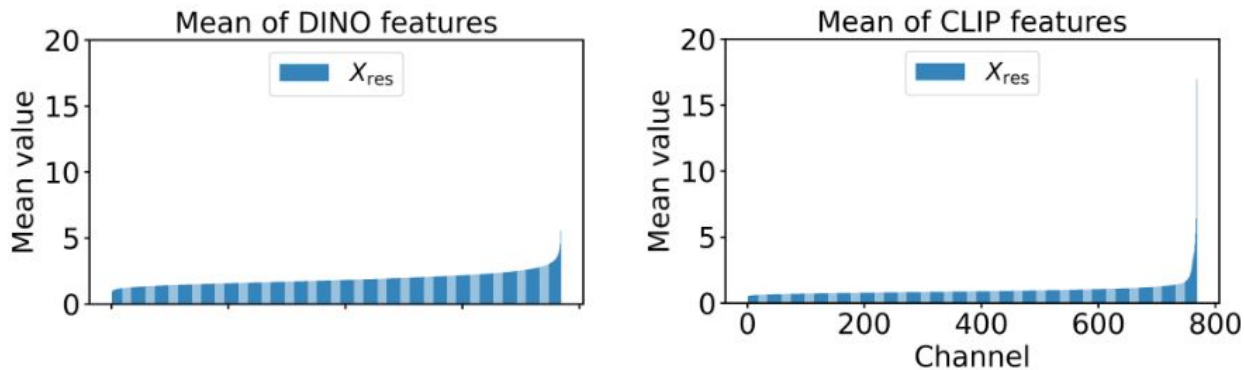
- 실험결과

- (3) DINO-B/16의 feature: 각 채널 별 X_{res} 평균값이 안정적으로 유지,
CLIP-B/16의 feature: 소수의 특정 채널이 X_{res} 평균값의 peak들을 지배함.

- > I.E., X_{res} 의 값 분포가 특정 채널에 몰려있다.

- > 각 feature들의 dominant channel이 같아, latent space에서 이 vector들의 방향이 유사하여, cosine similarity로 구분하기 어려움.

- > 이는 global 정보를 중요시하는 image recognition task에서는 괜찮지만, local 정보를 중요시하는 dense prediction task에서는 부적합.

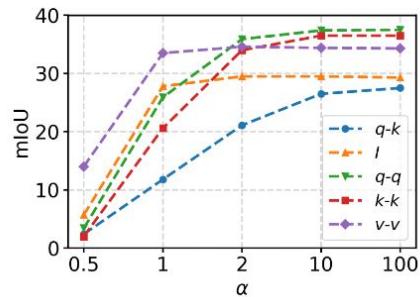


(c) Mean

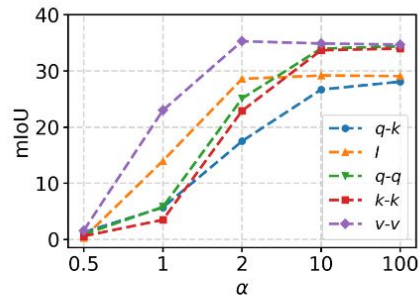
- Noise가 어떻게 발생하였는가 ?

- 실험결과
Residual connection의 영향을 줄일수록 성능 좋아진다.
- 결론
Residual connection 제거

$$X_{\text{sum}} = X_{\text{res}} + \alpha X_{\text{attn}}$$



(a) CLIP-B/16



(b) CLIP-L/14

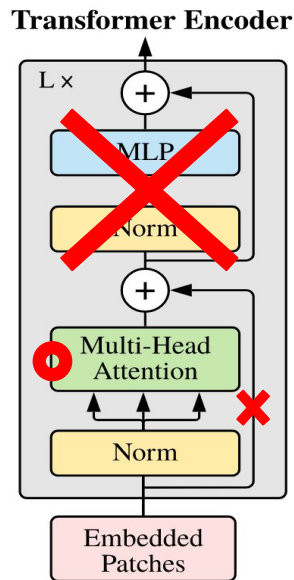
Fig. 6: Segmentation results w.r.t. the scaling factor α .

• Conclusion

1. Residual connection 제거
2. Feed-forward network 제거 (선행 연구 따라)
3. Query-query attention 적용 (선행 연구 따라)

$$X^{\text{visual}} = X_{\text{attn}} = \text{Proj}(\text{Attn}_{(\cdot)(\cdot)} \cdot v),$$

(6)



- Experiments

Table 1: Ablation results based on CLIP-B/16 architecture on five datasets *without* background class. RC denotes the residual connection.

| Attn | RC | FFN | VOC20 | Context59 | Stuff | Cityscapes | ADE20k | Avg. |
|-------|----|-----|-------------|-------------|-------------|-------------|-------------|-------------|
| $q-q$ | ✓ | ✓ | 68.4 | 24.9 | 14.7 | 20.8 | 7.6 | 27.3 |
| $q-q$ | ✓ | ✗ | 62.8 | 25.5 | 14.6 | 19.5 | 6.9 | 25.9 |
| $q-q$ | ✗ | ✓ | 77.6 | 31.8 | 21.0 | 23.4 | 14.7 | 33.7 |
| $q-q$ | ✗ | ✗ | 80.9 | 35.9 | 23.9 | 30.0 | 16.7 | 37.5 |

● Experiments

Table 2: Open-vocabulary semantic segmentation quantitative comparison on datasets *without* a background class. [†] denotes results directly cited from TCL [6]. SCLIP* denotes our reproduced results under the standard setting without class re-name tricks.

| Methods | Encoder | VOC20 | Context59 | Stuff | Cityscape | ADE20k | Avg. |
|----------------------------|----------|-------------|-------------|-------------|-------------|-------------|-------------|
| GroupViT [†] [44] | ViT-S/16 | 79.7 | 23.4 | 15.3 | 11.1 | 9.2 | 27.7 |
| CoCu [42] | ViT-S/16 | - | - | 13.6 | 15.0 | 11.1 | - |
| TCL [6] | ViT-B/16 | 77.5 | 30.3 | 19.6 | 23.1 | 14.9 | 33.1 |
| CLIP [35] | ViT-B/16 | 41.8 | 9.2 | 4.4 | 5.5 | 2.1 | 12.6 |
| MaskCLIP [†] [56] | ViT-B/16 | 74.9 | 26.4 | 16.4 | 12.6 | 9.8 | 28.0 |
| ReCo [†] [38] | ViT-B/16 | 57.7 | 22.3 | 14.8 | 21.1 | 11.2 | 25.4 |
| CLIPSurgery [26] | ViT-B/16 | - | - | 21.9 | 31.4 | - | - |
| SCLIP [40] | ViT-B/16 | 80.4 | 34.2 | 22.4 | 32.2 | 16.1 | 37.1 |
| SCLIP* [40] | ViT-B/16 | 78.2 | 33.0 | 21.1 | 29.1 | 14.6 | 35.2 |
| ClearCLIP | ViT-B/16 | 80.9 | 35.9 | 23.9 | 30.0 | 16.7 | 37.5 |
| CLIP [35] | ViT-L/14 | 15.8 | 4.5 | 2.4 | 2.9 | 1.2 | 5.4 |
| MaskCLIP [56] | ViT-L/14 | 30.1 | 12.6 | 8.9 | 10.1 | 6.9 | 13.7 |
| SCLIP [40] | ViT-L/14 | 60.3 | 20.5 | 13.1 | 17.0 | 7.1 | 23.6 |
| ClearCLIP | ViT-L/14 | 80.0 | 29.6 | 19.9 | 27.9 | 15.0 | 34.5 |

● Experiments

Table 3: Open-vocabulary semantic segmentation quantitative comparison on datasets *with* a background class. [†] denotes results directly cited from TCL [6]. SCLIP* denotes our reproduced results under the standard setting without class re-name tricks.

| Methods | Encoder | VOC21 | Context60 | Object | Avg. |
|----------------------------|----------|-------------|-------------|-------------|-------------|
| GroupViT [†] [44] | ViT-S/16 | 50.4 | 18.7 | 27.5 | 32.2 |
| SegCLIP [30] | ViT-S/16 | 52.6 | 24.7 | 26.5 | 34.6 |
| OVSegmentor [46] | ViT-B/16 | 53.8 | 20.4 | 25.1 | 33.1 |
| PGSeg [54] | ViT-S/16 | 53.2 | 23.8 | 28.7 | 35.2 |
| ViewCo [36] | ViT-S/16 | 52.4 | 23.0 | 23.5 | 33.0 |
| CoCu [42] | ViT-S/16 | 40.9 | 21.2 | 20.3 | 27.5 |
| TCL [6] | ViT-B/16 | 51.2 | 24.3 | 30.4 | 35.3 |
| CLIP [35] | ViT-B/16 | 16.2 | 7.7 | 5.5 | 9.8 |
| MaskCLIP [†] [56] | ViT-B/16 | 38.8 | 23.6 | 20.6 | 27.7 |
| ReCo [†] [38] | ViT-B/16 | 25.1 | 19.9 | 15.7 | 20.2 |
| CLIPSurgery [26] | ViT-B/16 | - | 29.3 | - | - |
| GEM [3] | ViT-B/16 | 46.2 | 32.6 | - | - |
| SCLIP [40] | ViT-B/16 | 59.1 | 30.4 | 30.5 | 40.0 |
| SCLIP* [40] | ViT-B/16 | 51.4 | 30.5 | 30.0 | 37.3 |
| ClearCLIP | ViT-B/16 | 51.8 | 32.6 | 33.0 | 39.1 |

● Experiments

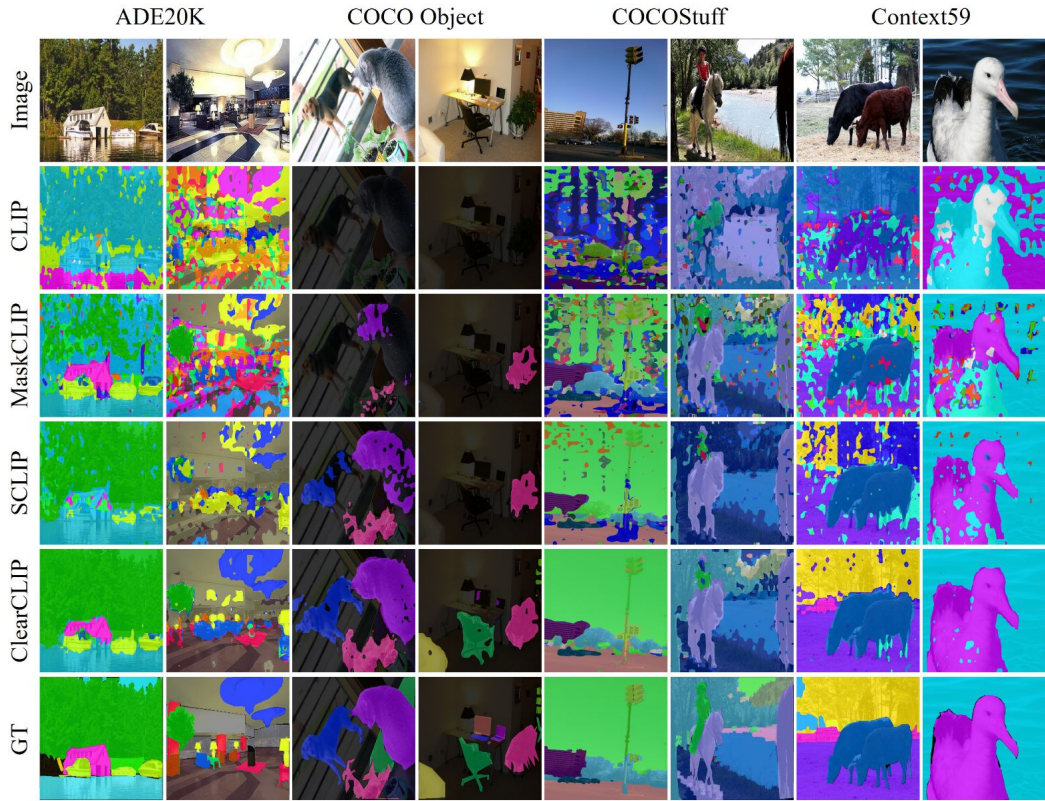


Fig. 7: Qualitative comparison between open-vocabulary segmentation methods.