ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference

Mengcheng Lan[1], Chaofeng Chen[1], Yiping Ke[2], Xinjiang Wang[3], Litong Feng[3⋆], and Wayne Zhang[3]

[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 layer에서 residual connection 제거

- **Semantic segmentation using CLIP**
  - ❑ Noisy 하다.
  - ❑ **이 noise는 어디서 왔으며, 어떻게 발생하였는가?**

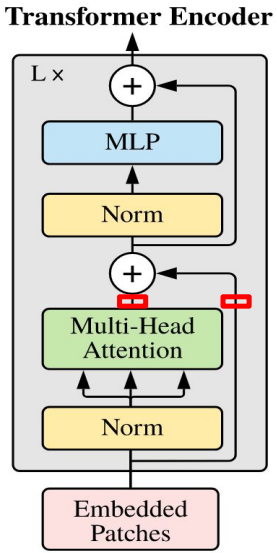
$$\mathcal{M} = \arg\max_{c} \cos(X^{\text{visual}}_{\text{dense}}, X^{\text{text}}). \quad (4)$$

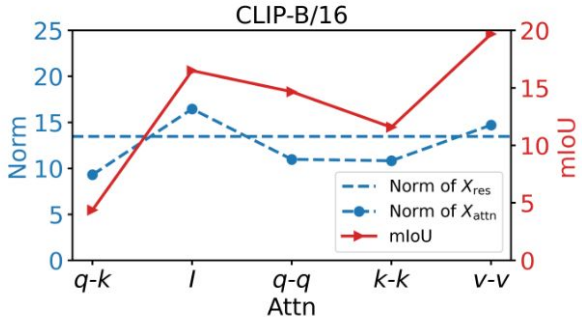
Image

CLIP

LAN, Mengcheng, et al. Clearclip: Decomposing clip representations for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 143-160.
ECCV 2024

- **Noise가 어디서 왔는가?**
  - ❑ 실험 분석
    (1) 성능과 attention output 크기는 positive correlation 관계.
    (2) CLIP-B/16: Residual norm이 작다 + Attention 수정이 성능 향상에 기여한다.
    (3) CLIP-L/14: Residual norm이 너무 크다 + Attention 수정이 효과 없는 경우가 생긴다.
  - ❑ 가설
    Residual connection이 CLIP의 segmentation 성능 향상 부진의 주 원인이다.





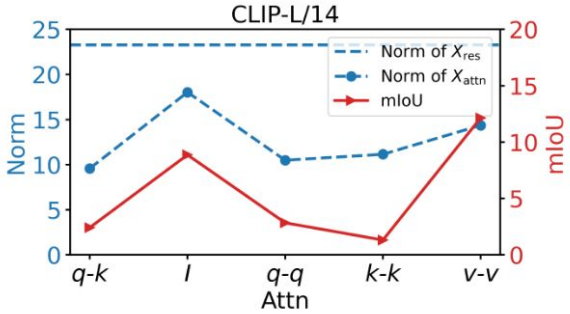
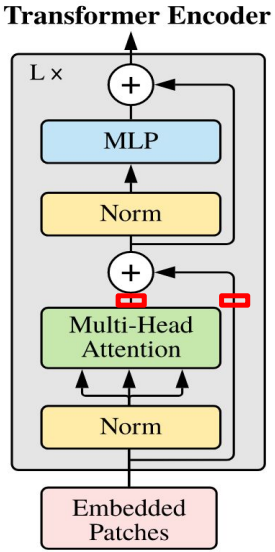
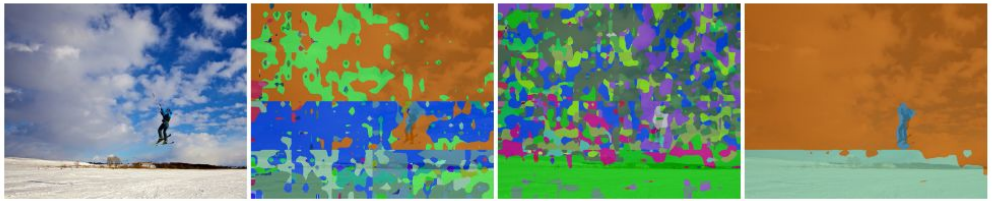


Fig. 2: Comparison of norms and mIoU of different attention mechanisms for CLIP-B/16 (left) and CLIP-L/14 (right). The norm curve of X_{attn} shows a positive correlation with the mIoU curve. A larger norm of X_{res} in CLIP-L/14 impedes the enhancement of performance through the revision of attention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

- **Noise가 어디서 왔는가?**
  - ❑ 가설 검증
    (1) Residual connection의 mIoU -> 0
    (2) Attention output의 mIoU > 최종 output의 mIoU
    (3) I.E., Residual connection은 image segmentation에 도움이 전혀 안되고 있다.
  - ❑ 결론
    CLIP을 사용했을때 segmentation map의 noise는 residual connection에서 비롯되었다.



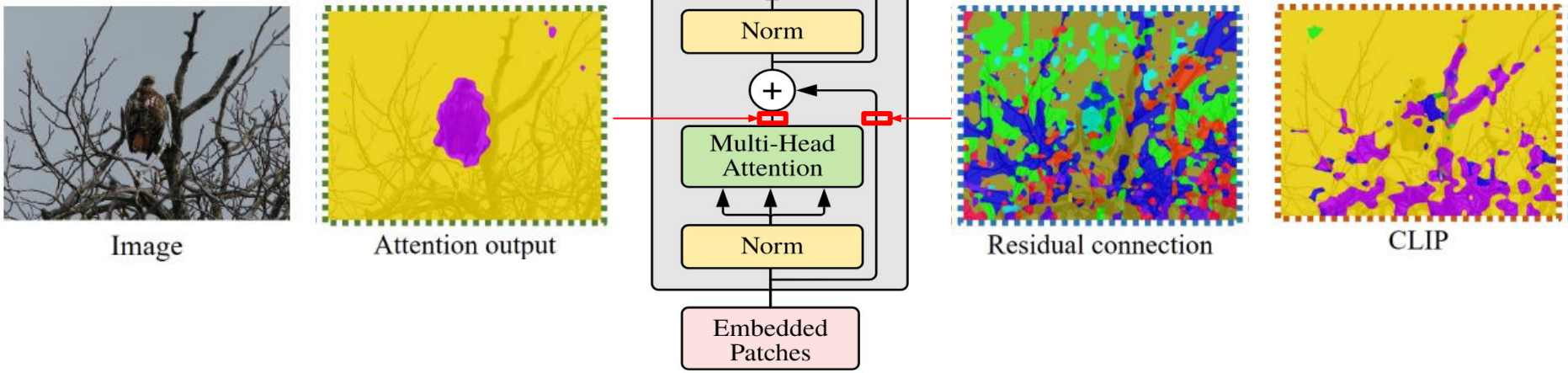


Features	mIoU
X_{sum}	4.4
X_{res}	0.01
X_{attn}	11.6

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).

LAN, Mengcheng, et al. Clearclip: Decomposing clip representations for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 143-160.

- **Noise가 어디서 왔는가?**
  - ❑ **Residual connection**



- **Noise가 어떻게 발생하였는가 ?**

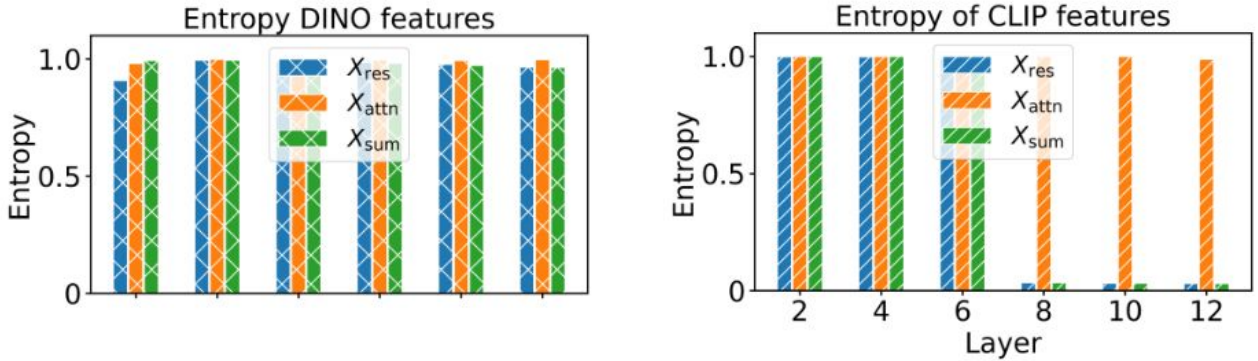
  ❑ **DINO vs CLIP entropy 비교**

  (1) DINO feature들은 entropy 일정하게 유지.

  (2) CLIP의 residual connection은 layer가 깊어짐에 따라 entropy가 급격히 감소.

  -> I.E., CLIP에서 layer가 깊어짐에 따라 일부 위치에만 peak값 집중.

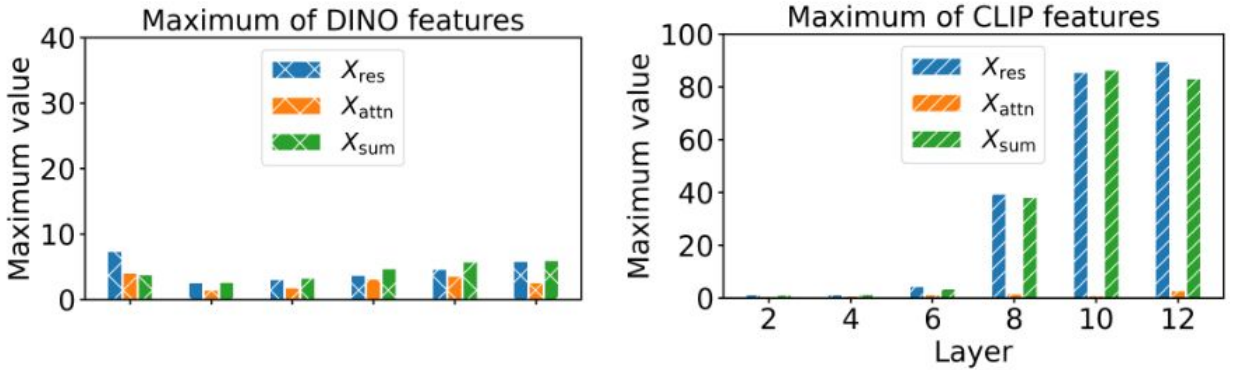
$$H(X^L) = -\frac{1}{\log(hw \times d)} \sum_{i,j} p(X^L_{i,j}) \log p(X^L_{i,j}), \quad p(X^L_{i,j}) = \frac{e^{X^L_{i,j}}}{\sum_{m,n} e^{X^L_{m,n}}}, \quad (5)$$



(a) Entropy

- **Noise가 어떻게 발생하였는가 ?**

  ❑ **DINO vs CLIP peak value 비교**
  (1) DINO feature들은 peak값이 10 이하로 안정적으로 유지.
  (2) CLIP의 residual connection은 layer가 깊어짐에 따라 peak값이 초기보다 90배 수준 증가.
  -> I.E., CLIP feature의 peak값 크기가 굉장히 크다.



(b) Maximum

LAN, Mengcheng, et al. Clearclip: Decomposing clip representations for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 143-160.
ECCV 2024

- **Noise가 어떻게 발생하였는가 ?**

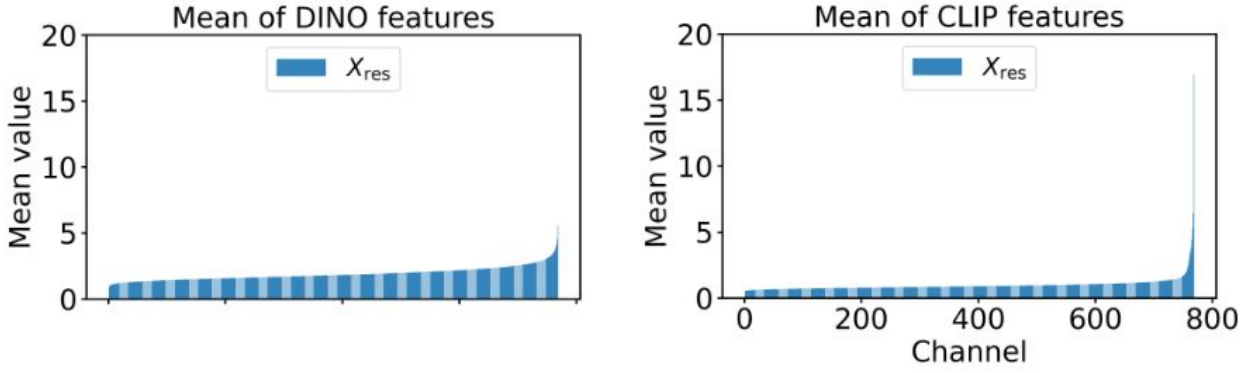
  ❑ **DINO vs CLIP channel-wise mean value 비교**

  (1) DINO의 residual feature 값은 일정하게 유지

  (2) CLIP의 residual feature는 특정 소수의 채널에 peak값들이 몰려있다.

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

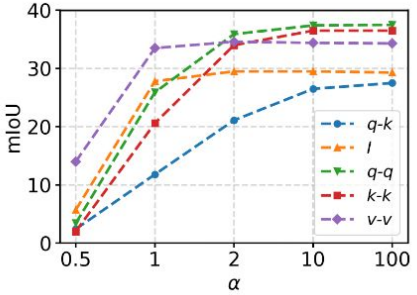
  -> 이는 global 정보를 중요시하는 image recognition task에서는 괜찮지만, local 정보를 중요시하는 dense prediction task에서는 부적합.



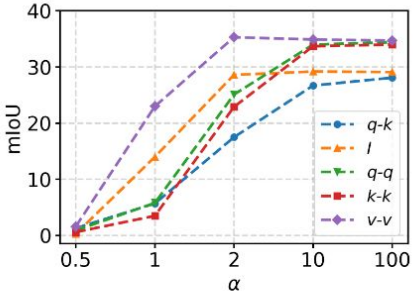
(c) Mean

- **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



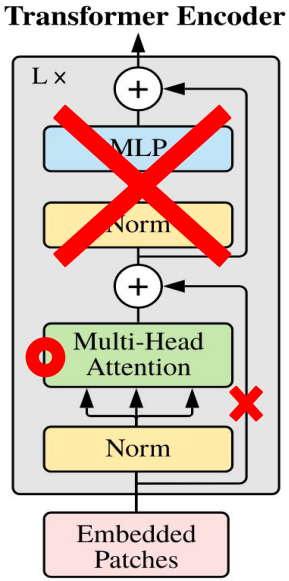
(b) CLIP-L/14

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

- **Conclusion**

1. Residual connection 제거
2. Feed-forward network 제거 [1, 2]
3. Query-query attention 적용 [3]

$$X^{\text{visual}} = X_{\text{attn}} = \text{Proj}(\text{Attn}_{(\cdot)(\cdot)} \cdot v), \tag{6}$$



[1] GANDELSMAN, Yossi; EFROS, Alexei A.; STEINHARDT, Jacob. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.
[2] LI, Yi, et al. A closer look at the explainability of Contrastive language-image pre-training. *Pattern Recognition*, 2025, 162: 111409.
[3] WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.

- **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**

LAN, Mengcheng, et al. Clearclip: Decomposing clip representations for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 143-160.
ECCV 2024

- **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**

LAN, Mengcheng, et al. Clearclip: Decomposing clip representations for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 143-160.
ECCV 2024

- **Experiments**

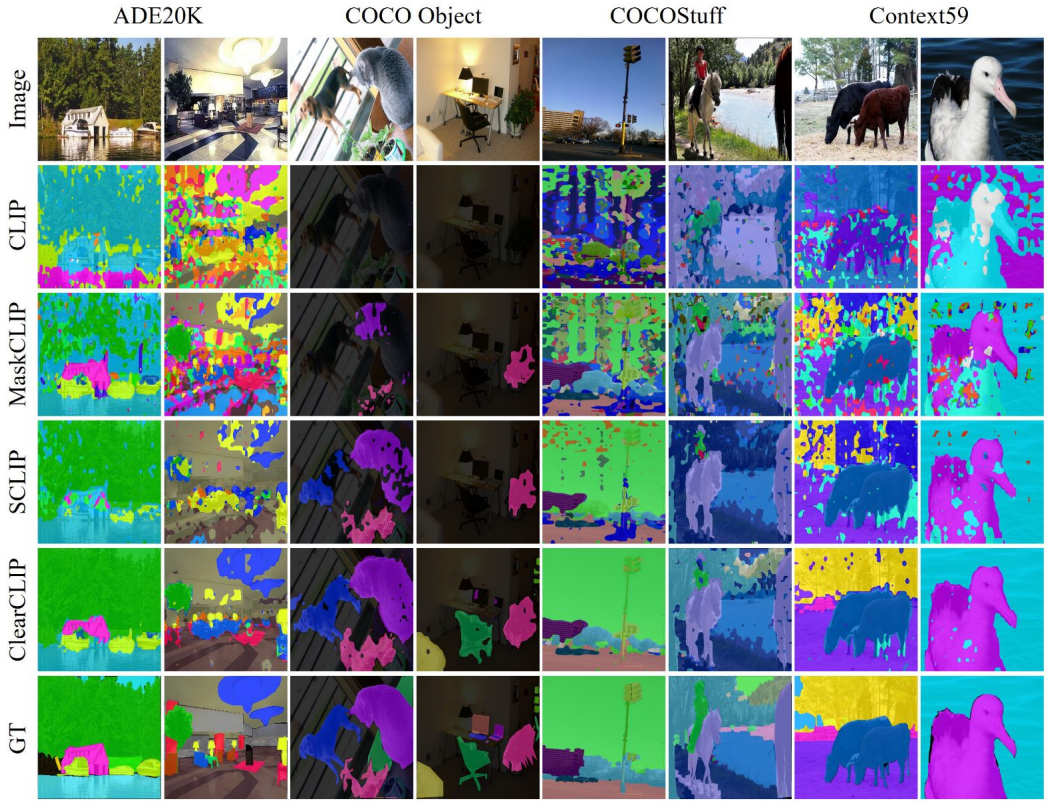


Fig. 7: Qualitative comparison between open-vocabulary segmentation methods.