Supplementary – 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

Appendix

A Ablation study with different backbones and datasets

We showcase the results of the ablation study for each dataset across different CLIP models in Fig. 1. It's clear that our method, which involves removing the residual connection and FFN, markedly enhances the open-vocabulary semantic segmentation capability of CLIP throughout all datasets. This enhancement is especially pronounced within the ViT-L/14 architecture, characterized by a larger norm of residual connection. These findings conclusively affirm the efficacy of our proposed methodology.

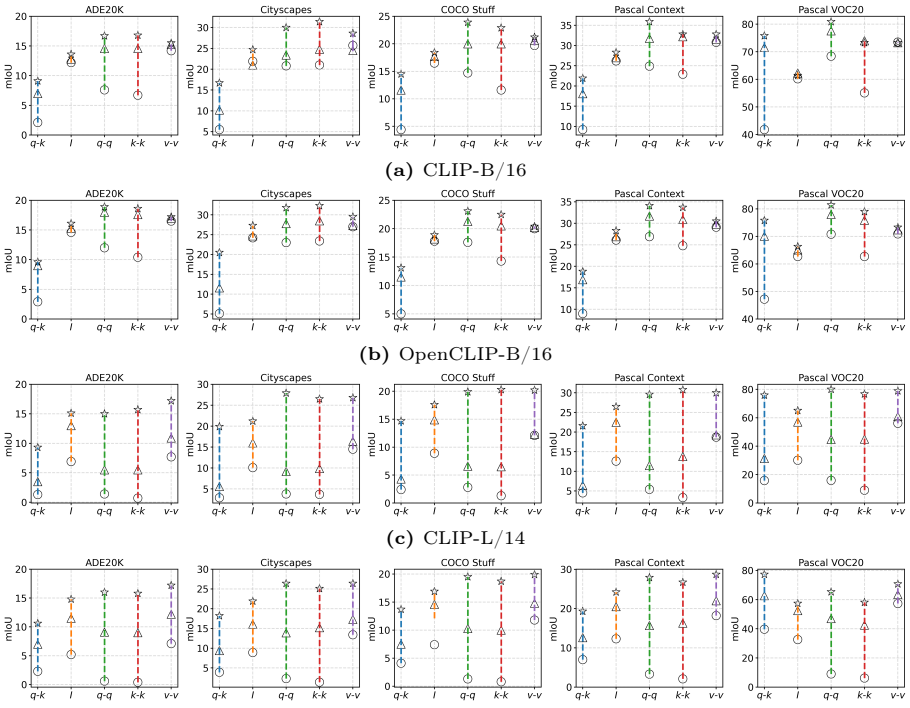
B Impact of channel-wise residual features

In this part, we investigate the effect of residual features with low intensity. Specifically, we conduct experiments by selectively reintroducing channels from residual features that have lower average values. We report the results of eliminating the top β high-value channels and the effect of normalizing X_{res} in Tab. 1. The best performance is achieved when $\beta \geq 70\%$. Additionally, normalizing X_{res} significantly reduces its scale, resulting in performance comparable to $\beta \geq 70\%$. These findings support our hypothesis that high-level supervision in CLIP emphasizes global feature direction in the residual latent space, which introduces noise into the residual features. For simplicity, we eliminate all channels in X_{res}.

C Integration across models

Our solution serves as a free lunch applicable to various architectures and segmentation models with ***just 2-3 lines of code modification***. Specifically, for

⋆ Corresponding author.



(a) CLIP-B/16

(b) OpenCLIP-B/16

(c) CLIP-L/14

(d) OpenCLIP-L/14

Fig. 1: Ablation study on each dataset under different architectures and attention mechanisms. ◯: original CLIP; △: CLIP w/o residual connection; ☆: CLIP w/o residual connection and FFN.

MaskCLIP and SCLIP, we achieve this by eliminating the residual connection and Feed-Forward Network (FFN) of the last self-attention layer. For GEM, we utilize the attention output from the final layer as the final representation. Importantly, we preserve the original attention mechanisms of these methods. For baseline models, *i.e.*, CLIP, BLIP, OpenCLIP, and MetaCLIP, we enhance them by incorporating our complete solution. The performance of different models on five datasets is summarized in Tab. 2. The results demonstrate that our solution consistently enhances the performance of existing models in open-vocabulary semantic segmentation tasks, showcasing its exceptional generalizability.

D Visualization of feature maps

To intuitively demonstrate how the residual connections affect the performance, we visualize the feature maps of X_{res}, X_{attn}, and X_{sum} for two randomly selected samples in Fig. 2. It is obvious that the X_{res} feature maps associated with the residual connections are characterized by peak values in one channel (highlighted in a red box), significantly surpassing the other channels. And X_{sum} is similar

Table 1: Average performance (mIoU) over all 8 datasets.

β (%)	0	5	10	30	50	70	100	Norm
Avg.	22.1	30.2	33.5	37.4	38.0	38.1	38.1	38.1

to X_{res}, indicating the big influence of X_{res} to the final feature. Conversely, the feature maps in X_{attn} demonstrate a more uniform distribution across channels. Given that the segmentation map is derived from the cosine similarity of feature vectors at each spatial location, such a disparity implies that the features in X_{sum} and X_{res} are less discernible compared to those in X_{attn}, thereby introducing noise into the segmentation results. This observation supports our proposal that *the high-level supervision in CLIP emphasizes the global feature direction in the residual latent space, making local feature vectors less distinguishable and leading to noise in residual features.*

E Additional qualitative examples

In this part, we present more qualitative results comparison between ClearCLIP and state-of-the-art methods. Figs. 3 and 4 show the results from COCOStuff, ADE20K and Pascal Context59 datasets respectively. Similar to the findings in the main text, the results of ClearCLIP exhibit much less noise than other methods, further underscoring the superiority of our method.

References

1. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding everything: Emerging localization properties in vision-language transformers. arXiv preprint arXiv:2312.00878 (2023)
2. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
3. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
5. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. arXiv preprint arXiv:2312.01597 (2023)
6. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023)
7. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)

Table 2: Average performance (mIoU) over 5 datasets without background class based on ViT- Base and Large architectures.

	VOC20	Context59	Stuff	Cityscape	ADE20K	Avg.
CLIP [4]	41.8	9.2	4.4	5.5	2.1	12.6
+ClearCLIP	80.9	35.9	23.9	30.0	16.7	37.5 +24.9
BLIP [3]	37.3	7.8	5.4	4.3	2.0	11.4
+ClearCLIP	73.5	31.4	21.3	23.8	13.5	32.7 +21.3
OpenCLIP [2]	47.2	9.0	5.0	5.1	2.9	13.8
+ClearCLIP	81.4	34.1	23.1	31.8	18.9	37.9 +24.1
MetaCLIP [6]	35.4	8.1	4.3	5.0	2.2	11.0
+ClearCLIP	78.3	34.8	23.5	27.9	17.4	36.4 +25.4
MaskCLIP [7]	74.9	26.4	16.4	12.6	9.8	28.0
+ClearCLIP	61.4	28.3	18.4	24.7	13.6	29.5 +1.8
SCLIP [5]	78.2	33.0	21.1	29.1	14.6	35.2
+ClearCLIP	77.9	35.6	23.6	31.0	17.0	37.9 +1.6
GEM [1]	79.9	35.9	23.7	30.8	15.7	37.2
+ClearCLIP	80.2	36.5	24.4	30.5	17.4	37.8 +0.6
CLIP [4]	15.8	4.5	2.4	2.9	1.2	5.4
+ClearCLIP	80.0	29.6	19.9	27.9	15.0	34.5 +29.1
BLIP [3]	22.5	5.8	2.4	3.8	1.5	7.2
+ClearCLIP	67.5	16.8	11.5	9.3	7.1	22.4 +15.2
OpenCLIP [2]	39.7	7.0	4.1	3.9	2.3	11.4
+ClearCLIP	65.3	27.9	19.5	26.4	16.0	31.0 +19.6
MetaCLIP [6]	22.7	6.2	3.6	5.1	2.2	8.0
+ClearCLIP	78.2	30.3	20.5	25.6	16.4	34.2 +26.2
MaskCLIP [7]	30.1	12.6	8.9	10.1	6.9	13.7
+ClearCLIP	65.1	26.5	17.6	21.2	15.1	29.1 +11.1
SCLIP [5]	60.3	20.5	13.1	17.0	7.1	23.6
+ClearCLIP	79.2	30.6	20.5	27.8	15.6	34.7 +15.4
GEM [1]	80.3	26.4	17.6	22.6	11.6	31.7
+ClearCLIP	79.7	29.9	19.4	25.9	14.2	33.8 +2.1

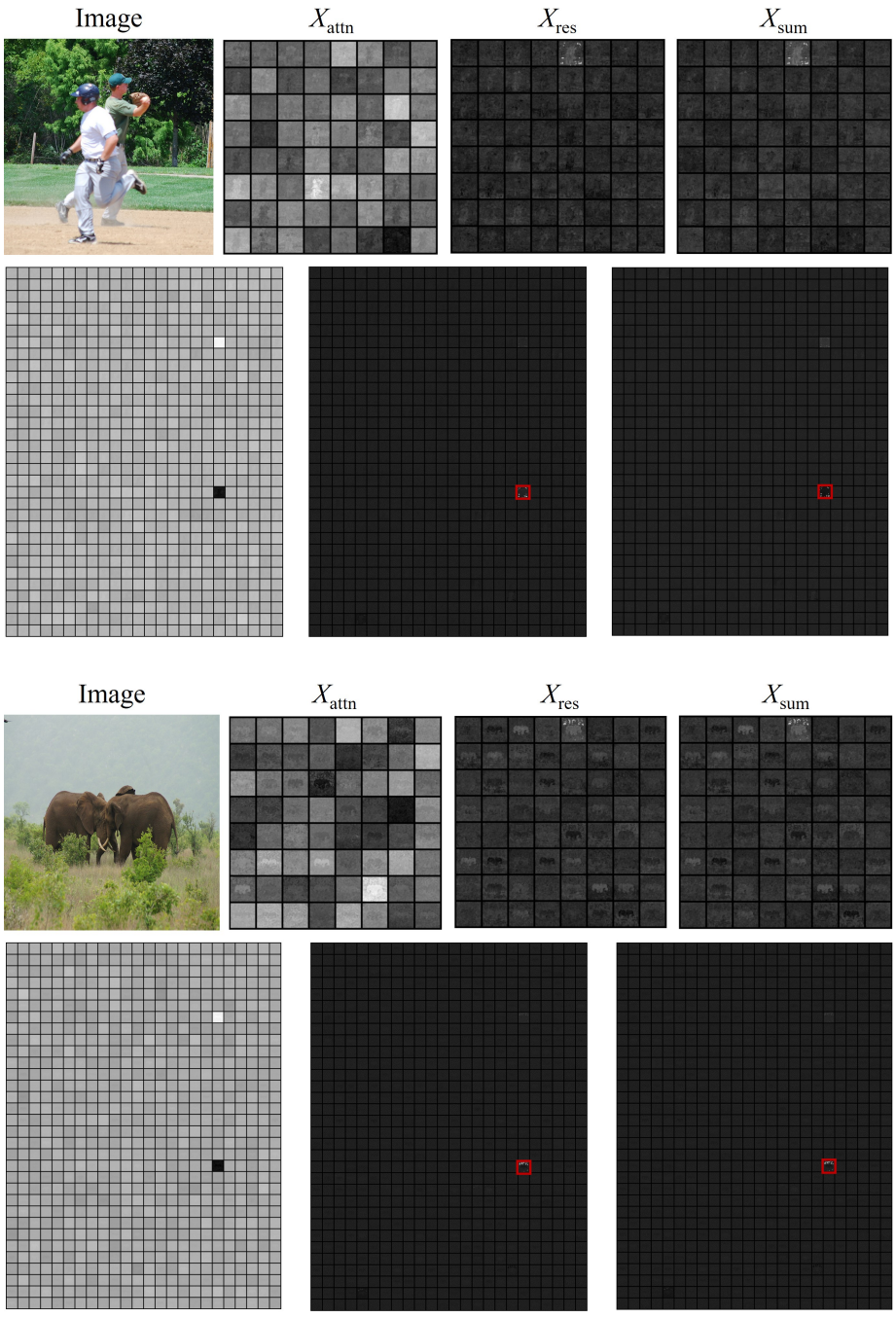
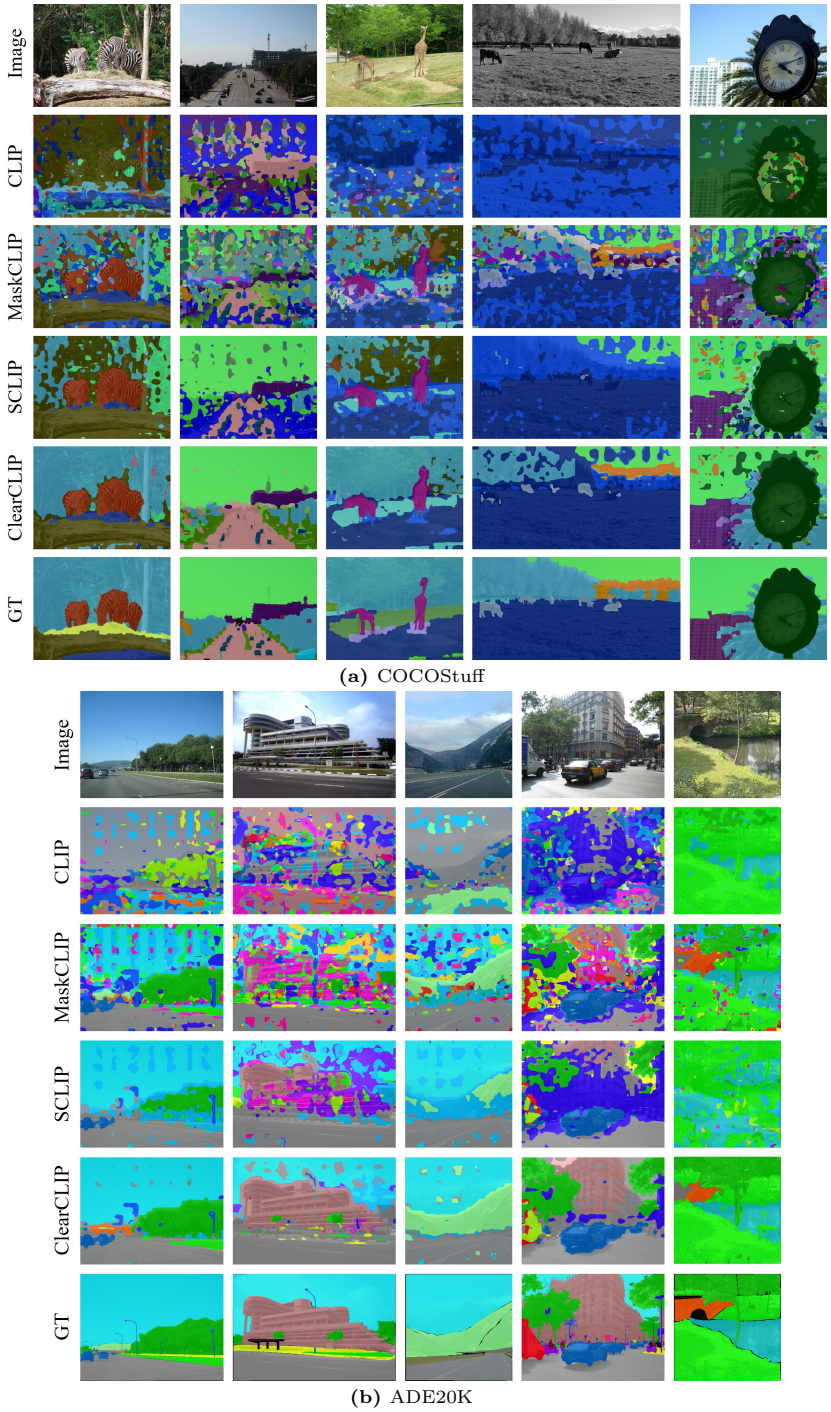


Fig. 2: Visualization of feature maps with CLIP for two randomly selected examples from the COCOStuff dataset. The first row shows the first 64 feature maps of each type, while the second row displays all 768 feature maps of each type.



(a) COCOStuff

(b) ADE20K

Fig. 3: Qualitative comparison between different open-vocabulary segmentation methods on (a) COCOStuff and (b) ADE20K datasets.

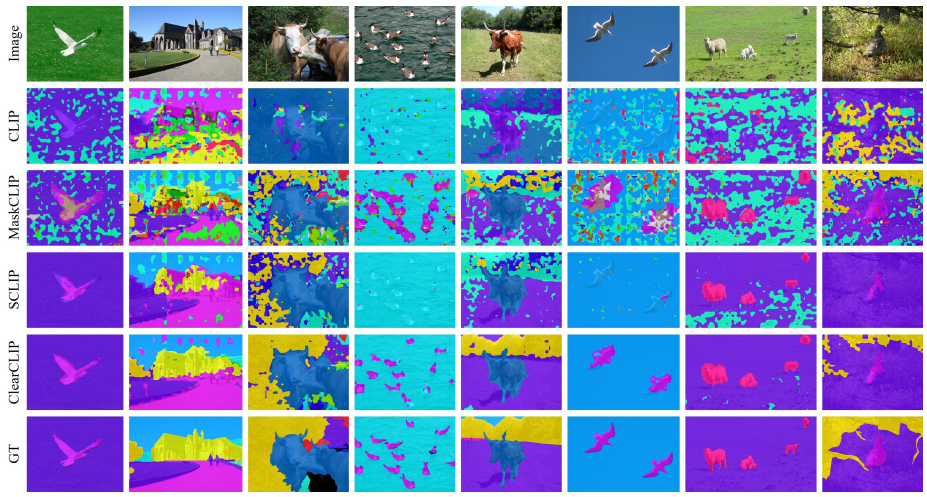


Fig. 4: Qualitative comparison between different open-vocabulary segmentation methods on the Pascal Context59 dataset.