Supplementary Materials: Explore the Potential of CLIP for Training-Free Open Vocabulary Semantic Segmentation

Overview

This is the supplementary file for our submission titled *Explore the Potential of CLIP for Training-Free Open Vocabulary Semantic Segmentation*. This material supplements the main paper with the following content:

A Details of Our Method CLIPtrase

This section primarily supplements some method details and consists of the following three parts: 1). further analysis of global patches, 2). efficient variants for effective semantic correlation recovery calculation, and 3). reasons for using mask classification.

A.1 More Explanation of Global Patch

Observation. As mentioned in the main text, the [CLS] token does not directly weight all patches of the entire image through an attention weight map to obtain the final result as we might imagine. Instead, the [CLS] token interacts with the global patches to serve as a stepping stone for completing the process. Global patches: When the [CLS] token extracts overall information from the image,

there are several global patches that have high attention weights for any local patch in the image, providing a global view. The [CLS] token does not directly interact with all patches but focuses on these global patches to obtain image-level information, while disregarding the weights of other local patches.

Furthermore, through our observations, we find that this phenomenon is commonly observed from the 6-th to the 11-th layers of the CLIP ViT, while with some alleviation in the final layer strangely.

Analysis. Since the CLIP model is trained solely on a contrastive learning objective function and there are no additional constraints on the [CLS] token, we believe that this phenomenon is spontaneously formed by the model during the training process. We explore why CLIP spontaneously forms global patches by considering the benefits they provide to the CLIP. By starting from the advantages offered by such global patches, we can gain insights into the reasons behind their spontaneous formation in the model.

We think there are two main reasons:
First of all, one of the reasons is the redundancy of image information. From the perspective of information content, even if they describe the same object, the information content of images is much higher than that of text. And in the image, in addition to the main object described, most areas are other objects and background, which are redundant for text category features. This aspect causes [CLS] token to face a lot of redundant information when extracting image-level feature representation.

The emergence of global patches can effectively alleviate this phenomenon. The global patch itself can act as a filter, allowing the [CLS] token to learn features that better match the texts.

This can be seen from the semantic correlation visualization of images in Section D. Often, for images with large-area backgrounds, such as large-area sky, ocean, etc., the more global patches there are. This is similar to the conclusion in the StreamingLLM [51] model: Xiao et al. believe that the first token in GPT acts as a "trash can", causing redundant information to be placed in this token. Our global patch here also plays a similar role and helps filter redundant information.

The second point is due to the training trend of the model. Just like general models tend to learn sparse features, CLIP's [CLS] token also has this tendency when faced with complex patch tokens. A large number of patch tokens increases the difficulty of learning of [CLS] token. The emergence of global patches can make the patch interaction process also show a sparse trend, and [CLS] token only interacts with global patches. This greatly reduces the learning difficulty of [CLS] token.

Effects. Although the emergence of this global patch has several benefits we mentioned above, it also greatly destroys the semantic correlation between local patches.

Global patches maintain high weighted attention with all local patches. However, due to the existence of softmax, the high weights with the global patch

gradually squeeze the correlation weights between the original patches, causing the local patches to lose their ability to pay attention to adjacent or same semantic patches.

In the final layer of CLIP-ViT, due to the objective function aligning the [CLS] token with the text, certain semantic information is fed back from the [CLS] token to the patches across the entire image, leading to a partial recovery of semantic relevance among local patches. However, due to the influence of the global patches mentioned earlier, this recovery is limited and cannot fully counteract the suppression of original semantic relevance by the global patches in the attention matrix.

Therefore, the presence of global patches leads to a lack of semantic relevance among local patches, which is the main reason why CLIP is not well-suited for dense feature tasks such as semantic segmentation.

A.2 Semantic Correlation Computing

In the original computation of semantic correlation recovery, we perform self-correlation separately for each branch and each head. We then calculate the mean to obtain a more stable measure of semantic relevance:

$$\boldsymbol{q}_i = \sigma_{i,q}(\boldsymbol{x}),\ \boldsymbol{k}_i = \sigma_{i,k}(\boldsymbol{x}),\ \boldsymbol{v}_i = \sigma_{i,v}(\boldsymbol{x}),\ i \in [0, H-1] \tag{1}$$

$$\boldsymbol{w}_{i,q} = \gamma(\boldsymbol{q}_i, \boldsymbol{q}_i),\ \boldsymbol{w}_{i,k} = \gamma(\boldsymbol{k}_i, \boldsymbol{k}_i), \boldsymbol{w}_{i,v} = \gamma(\boldsymbol{v}_i, \boldsymbol{v}_i) \tag{2}$$

$$\boldsymbol{w} = \frac{1}{3H} \sum_{i=0}^{H-1} (\boldsymbol{w}_{i,q} + \boldsymbol{w}_{i,k} + \boldsymbol{w}_{i,v}) \tag{3}$$

where σ, γ is the linear layer and semantic correlation we mentioned in the main text, respectively. However, performing separate computations on each head can reduce computational efficiency. Therefore, a simple approach to streamline this calculation process is to concatenate the outputs from each head:

$$\boldsymbol{q}_i = \sigma_{i,q}(\boldsymbol{x}),\ \boldsymbol{k}_i = \sigma_{i,k}(\boldsymbol{x}),\ \boldsymbol{v}_i = \sigma_{i,v}(\boldsymbol{x}),\ i \in [0, H-1] \tag{4}$$

$$q = \phi(\boldsymbol{q}_0, ..., \boldsymbol{q}_{H-1}),\ k = \phi(\boldsymbol{k}_0, ..., \boldsymbol{k}_{H-1}),\ v = \phi(\boldsymbol{v}_0, ..., \boldsymbol{v}_{H-1}) \tag{5}$$

$$w = \frac{1}{3}(\gamma(q, q) + \gamma(k, k) + \gamma(v, v)) \tag{6}$$

where ϕ is the operation of concatenation. Such a replacement solution reduces the computational burden and has almost no impact on performance. However, when explaining ideas, we prefer to use the calculation method in the main text because it is more intuitive.

Table 1: SCR rationality validation.

input=224, mIoU	VOC	ADE	COCO	PC	Avg.
w/o. SCR	67.77	3.01	5.40	10.85	21.76
$< x_i, x_i >$	62.69	11.55	16.40	25.11	28.94
$< x_i, x_j >$	75.61	15.58	21.39	32.13	36.18
$COS(x_i, x_j)$	77.58	15.79	21.82	32.55	36.94
$COS(x_i, x_j), j \in GT$	77.71	16.50	22.57	33.89	37.67
CLIPtrase (inner product SCR)	63.55	7.49	12.64	19.94	25.91
CLIPtrase (Cosine SCR)(Ours)	80.95	16.35	22.84	33.83	**38.50**

Rationale of semantic correlation recovery. In Table 1, we gradually refine the scope of attention to estimate an increase in the degree of semantic correlation restoration (SCR): 1): without SCR; 2): self attention: focus on own patches, 3): normal attention: by inner product or Cosine, 4):attention within ground truth regions. It is evident that global patches significantly disrupt semantic correlation from Figure 1,2,4 of the paper. As the region of SCR becomes more precise, the segmentation performance gradually improves in Table 1, which indicates a diminishing influence of global patch. *We demonstrate that as the degree of SCR increases, the segmentation performance improves. This serves as the proof that global patch is one of the reasons causing limited segmentation.* Furthermore, both Cosine similarity and inner product yield similar results from row 3,4 in the Table 1. However, the unbounded range from inner product has a negative impact (lines 209-215, 387-392 in the paper) on subsequent clustering module, which shown in last two rows of Table 1. Therefore, we select the Cosine similarity to implement the SCR.

A.3 Mask Classification

In the main paper, our clustering design is mainly to adaptively obtain the mask corresponding to each object, and then let all pixels in the mask jointly determine the category of this area. So there may be a question here: *why not do patch-text prediction directly?*

We explain in detail the reasons for this operation:

First, such a clustering operation can form masks of the same semantic areas through the common features of the semantic similarity matrix. This idea of mask and joint decision-making make most of the correct predictions in this semantic area to improve some noise predictions. Through the performance and ablation experiments in the main paper we are able to prove the effectiveness of this approach and can effectively improve the prediction noise in the image.

In addition, this approach is closer to the current methods that requires training based on MaskFormer [7], which lays the foundation for us to apply it to the models those require training.

In fact, a more consistent approach with mask-based training models such as [19, 53] is to use the mask obtained by clustering to perform mask pooling

Table 2: Our approach combined with MaskCLIP or DINOv2.

input=224, mIoU	VOC	ADE	COCO	PC	Avg.
Ours + DINOv2	80.23	15.79	21.82	32.55	37.59
CLIPtrase(Ours)	80.95	16.35	22.84	33.83	**38.50**

on the CLIP features, and use the masks as the attention bias to generate its corresponding [CLS] token for each mask. However, since this method requires separately designing and training some layers for attention bias, which violates our original intention of directly adapting CLIP to the semantic segmentation task, it was abandoned.

In addition, to further demonstrate the effectiveness of our self-correlation and the rationale behind using masks, we compare the results obtained by using DINO [2] for masking. Table 2 presents a detailed performance comparison, and it can be observed that our method exhibits similar performance to the DINO model during the masking process after undergoing self-correlation recovery.

B More Experiments

B.1 Experiment Details

In this section, we present further details and configurations utilized in our experiments.

Environment. The environment we use is: CUDA version: 11.3, PyTorch: 1.12.1, GPU: NVIDIA RTX 3090*1, CLIP: CLIP-B/16, local implementation.

Data Proprocessing. The data preprocessing and data enhancement solutions in this paper are consistent with CLIP preprocessing and do not add any additional operations. We maintain the order of operations of resize, crop, and normalize. The mean and variance of normalizing on the image are:[0.48, 0.46, 0.41], [0.27, 0.26, 0.28] (two decimal places).

Hyper-parameters. Our image sizes in experiments are 224, 336. In the subsequent clustering of the attention weights, $\epsilon = 0.7$, $min\_sample = 3$. The text prompts are consistent with the official implementation of CLIP. The average of 80 short sentence prompts is taken to represent the final text feature.

Datasets. We have a total of 9 benchmarks in the experiments, involving 4 datasets:

- **COCO** [1]: There are a total of 80 object classes and 91 stuff classes. We use a total of 171 classes as COCO-stuff, and a total of 81 classes of objects and additional background classes as COCO-object.

Table 3: Performance comparison of four evaluation indicators for image size 224. The best average performance under each metric is bolded.

Image size=224	CLIP				SCLIP				CLIPtrase (Ours)			
Evaluation	pAcc	mAcc	fwIoU	mIoU	pAcc	mAcc	fwIoU	mIoU	pAcc	mAcc	fwIoU	mIoU
COCO-obj	23.45	20.58	14.39	12.98	49.05	59.43	37.38	40.68	50.08	62.50	38.19	43.56
VOC21	42.21	47.21	30.47	17.54	77.52	83.27	66.46	49.54	78.63	84.11	67.67	50.88
PC60	21.84	21.22	11.02	8.80	49.24	51.55	35.04	28.09	52.14	56.08	37.61	29.87
COCO-stuff	10.85	12.02	5.66	4.68	35.24	40.02	24.42	21.00	38.90	44.47	26.87	22.84
VOC20	58.50	58.35	44.02	41.88	87.51	89.92	79.21	77.58	89.68	91.40	82.49	80.95
PC59	24.60	21.78	13.29	9.70	55.69	52.58	42.13	31.58	58.94	57.08	45.28	33.83
PC459	16.94	4.02	9.31	1.92	41.15	18.63	32.72	8.48	44.18	21.53	35.22	9.36
ADE150	5.20	7.61	2.59	2.25	33.00	34.43	23.44	14.46	38.57	39.17	27.96	16.35
ADEfull	2.69	3.13	1.21	0.76	20.06	16.46	14.55	5.43	25.45	18.78	18.99	6.31
AVG.	22.92	21.77	14.66	11.17	49.83	49.59	39.48	30.76	**52.95**	**52.79**	**42.25**	**32.66**

- **PASCAL CONTEXT** [33]: There are 459 categories in total, and we select 59 common categories as PC59, and all categories as PC459. In addition, we add an additional background as PC60 based on PC59.
- **PASCAL VOC2012** [16]: There are 20 categories in total, which we refer to as VOC20. In addition, we add common background categories as VOC21 based on VOC20.
- **ADE20K** [59]: There are 847 classes in total, and we select 150 of them as ADE150, and all classes as ADEfull.

Evaluation Protocol. Following the common practice [7, 17, 54], we use the mean of class-wise intersection over union (mIoU) to measure the performance. In addition, we also report on the performance of mean accuracy (mAcc), pixel accuracy (pAcc), and frequency weighted intersection over union (fwIoU) to comprehensively verify the performance of our method from multiple aspects.

B.2 More Experiment Results

We mainly reproduce the semantic segmentation of CLIP [36] and SCLIP [48] on various datasets, and mainly compare the two models to analyze the effectiveness of our model in the training-free open vocabulary semantic segmentation.

Compared with the results in the main text, we mainly supplement the different image sizes of pAcc, mAcc, fwIoU and mIoU using CLIP, SCLIP and our own methods on each dataset in Table 3 and 4, to prove the effectiveness of our method from a more comprehensive perspective. Judging from the improvement of pAcc, our method distinguishes and clusters objects with different semantics in the image, rather than just focusing on the main object. Although some objects are uniformly regarded as background, we think this advantage will have greater potential in subsequent downstream tasks.

Table 4: Performance comparison of four evaluation indicators for image size 336. The best average performance under each metric is bolded.

Image size=336	CLIP				SCLIP				CLIPtrase (Ours)			
Evaluation	pAcc	mAcc	fwIoU	mIoU	pAcc	mAcc	fwIoU	mIoU	pAcc	mAcc	fwIoU	mIoU
COCO-obj	22.53	19.85	13.55	12.63	48.34	57.62	36.79	40.43	50.01	62.55	38.24	44.84
VOC21	43.16	45.97	31.39	17.31	79.41	82.66	68.64	51.79	79.93	85.24	69.10	53.04
PC60	21.84	21.25	11.20	8.91	49.95	51.45	35.88	29.00	53.21	56.43	38.76	30.79
COCO-stuff	11.11	11.95	5.82	4.70	35.97	39.51	25.08	21.61	40.14	45.09	27.96	24.06
VOC20	57.12	57.92	42.60	41.06	87.03	90.97	78.48	79.12	89.51	91.77	82.15	81.20
PC59	24.53	21.81	13.42	9.82	56.50	52.49	43.08	32.59	60.15	57.47	46.64	34.92
PC459	16.94	3.94	9.50	1.88	42.07	18.43	33.73	8.97	45.77	20.72	36.67	10.11
ADE150	5.64	7.65	2.89	2.3	33.84	32.28	24.27	14.76	39.92	37.75	29.17	17.04
ADEfull	2.96	3.04	1.39	0.79	21.33	15.3	15.9	5.28	26.73	17.99	20.3	5.89
AVG.	22.87	21.49	14.64	11.04	50.49	48.97	40.21	31.51	**53.93**	**52.78**	**43.22**	**33.54**

Table 5: Results on unsupervised semantic segmentation. We use datasets that are consistent with the baselines, with the dataset suffix indicating the number of categories.

mIoU	PiCIE [8] (CVPR'21)	STEGO [23] (ICLR'22)	HP [38] (CVPR'23)	SmooSeg [25] (NIPS'23)	**Ours**
coco27	13.8	24.5	24.6	26.7	**30.8**
cityscape27	12.3	21.0	18.4	18.4	**20.0**

In addition, under these measurement standards, no matter with which resolution, our model is about 3% higher than SCLIP, which can more comprehensively prove the effectiveness of our method.

B.3 More applications

In addition to the application mentioned in the main paper that involve combining our model with SAM, it can be applied to many other areas as well.

Combining with unsupervised semantic segmentation. Our CLIPtrase model, by recovering the internal local correlations within CLIP through self-correlation, enables CLIP to provide more semantic contextual details within images. This kind of information is invaluable for tasks that lack pixel-level annotations, such as semi-supervised and unsupervised semantic segmentation.

Taking unsupervised semantic segmentation tasks as an example, we combine the generalization capability of Clip with the region correlations recovered by our method and apply them to the unsupervised task. We compare our approach with SOTA models in Table 5 and find that our CLIPtrase plays a significant role in unsupervised semantic segmentation.

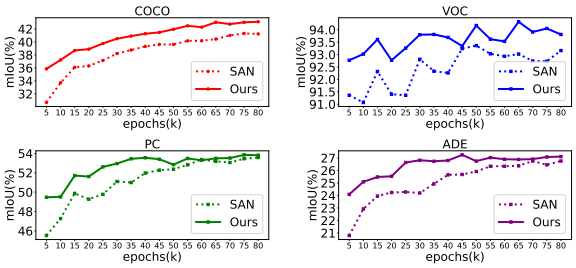

Fig. 1: The convergence of our approach combined with SAN.

Combining with training OVSS. Additionally, our method can also assist OVSS models that require training. For instance, for the SAN [53] model, we utilize the features and masks from CLIPtrase and perform mask average pooling (MAP) to initialize the query embedding in SAN.

Our method helps improve the training speed of the SAN model. SAN itself is a lightweight model, when combined with our approach, which shown in Figure 1, it further reduces the training burden, resulting in a doubling of the convergence speed on specific datasets.

C Related Work

C.1 Contrastive Language-Image Pre-training

Contrastive Language-Image Pre-training (CLIP) [36] is a large multi-modal foundation model, which utilizes the contrastive training of aligning visual and text corresponding category features, greatly improves the generalization on unseen samples. Currently CLIP is widely used in Few-Shot/Zero-Shot Learning (FSL/ZSL) [22, 26, 29, 61, 62], Prompt learning [22, 26, 61, 62] and Out-of-Distribution (OoD) [40] tasks. Later, researchers begin to apply CLIP to dense feature tasks [39, 49, 57, 58] such as semantic segmentation [29, 41].

Li et al. [27] elaborate on the inherent noise problem of CLIP and introduce it into the open vocabulary task from the perspective of explainability through self-attention improvement. Unlike pipelines that generally fine-tune pre-trained models on additional data sets, the CLIP encoder often needs to be frozen and cannot be fine-tuned because it needs to maintain alignment with the text feature space [60]. Therefore, researchers currently prefer to use clip directly as an encoder to obtain preliminary features to inherit its excellent generalization ability, and pay more attention to design sophisticated decoders [9, 13, 18, 37, 50] to refine the image-level features to adapt to dense feature tasks.

C.2 Open-Vocabulary Semantic Segmentation

Open-vocabulary semantic segmentation extends segmentation [5, 10–12, 20, 24, 32, 35, 45–47] and refers to segment semantic regions via textual names or descriptions for the open world without any mask annotations. Early works [60]

verify the importance of modal alignment in CLIP, and common downstream fine-tuning may destroys its generalization ability. MaskCLIP [60] attempts to improve the Vision Transformer (ViT) [15] structure of CLIP to allow the model to obtain coarse feature localization, and combines transductive learning to improve performance. CLIP-Surgery analyzes the difficulty of the current semantic segmentation task introduced by CLIP from the perspective of image-text noise, and made certain improvements to the model using the idea of self-attention. SCLIP [48] inherits the idea of self-attention of MaskCLIP and directly adapts the improved CLIP structure to the semantic segmentation task.

Both CLIP-Surgery and SCLIP utilize the idea of self-attention to improve CLIP, while only CLIP-Surgery mentions the noise problem caused by the open category of text. None of them explore and analyze why CLIP lacks the semantic correlation between patches. Our work complements this point that it is the global patch formed during the attention interaction between [CLS] token and patches that leads to this.

The above methods attempt to directly apply CLIP to semantic segmentation tasks. In the methods of training on additional segmentation datasets, CLIP tends to be used as an encoder. We roughly divide them into two ideas:

Decoder-based. Inspired by MaskFormer [7] and Mask2Former [6], open vocabulary semantic segmentation widely use the pipeline of mask generation + mask classification. This method trains the refined features using a pixel decoder and utilizes an additional query decoder to aggregate the refined features at different positions. It employs query embedding to obtain masks for different objects, calculates the similarity between query embedding and texts, and then weights the query masks accordingly to ultimately yield boundaries and categories for each class.

Thanks to the excellent architecture of MaskFormer, the effect of the mask generation module masks great progress. However, The generalization performance of the above mask classification on unseen samples is always the bottleneck of this problem [28] due to the limitation of training scale. Therefore, some researchers combine it with CLIP to complete the classification of masks through the generalization of CLIP and improve the model performance [14,19,31,42–44].

For example, Xu et al. [53] design the side network to generate masks in parallel with CLIP, and then use the masks as the attention bias to learn the corresponding [CLS] token for each mask and complete mask classification. Similar ideas include TCL [3], which enables CLIP simultaneous participation in the mask generation and classification stages by reusing the visual branch.

Another method with the idea of combining CLIP and masks is GroupViT [4, 52]. It designs multiple group tokens, continuously aggregates them during the text guidance training process, and finally makes each group token cover the area of a specific object. However, compared with the above idea, the prediction results of this method often contain messy and tiny segmented areas.

Fine-tuning. In addition, there are also methods that advocate direct fine-tuning of the CLIP [28, 30, 34, 47, 55, 56], among which the typical method is

OV-Seg [28]. It believes that the classification of images after masks is may have the domain shift problem, so OV-Seg use additional masked dataset to fine-tune the CLIP, adapt to the special need of mask classification. MAFT [21] draws on the idea of SAN [53] and fine-tunes the process of generating corresponding [CLS] tokens based on the attention bias formed by different masks, so that it can also achieve the purpose of improving mask classification performance.

There are also methods to directly use the contrast learning idea of CLIP to retrain the encoder at the patch or pixel level. For example, PACL [34] refines this alignment to the patch level and improves semantic coherence issues.

Despite the decent results demonstrated by the two approaches, they still possess their respective issues. For the decoder-based method, the CLIP's features are obtained via a frozen CLIP model without specific adjustments to adapt them for semantic segmentation. Instead, this issue is addressed through the addition of an external decoder. On the other hand, although fine-tuning allows for a certain degree of adaptation of CLIP's image-text aligned features, it carries the risk of overfitting to specific scenarios, e.g., the domain of the masked images used for fine-tuning, leading to a decline in performance. Therefore, we need to consider whether it is possible to optimize CLIP's features without finetuning, in order to unearth more information that can aid in semantic segmentation. To answer this question, we start by investigating the correlations between the [CLS] token and the patch tokens.

D Visualization

In this section, we mainly visualize the effects of several modules in the method in detail.

Figure 2 illustrates the global patch problem we mentioned above, and the results improved with our semantic relevance recovery method. We demonstrate this global patch phenomenon and the performance of our improved method in various image situations through richer visualizations.

In the original CLIP, the response heat map of the area we select is completely inconsistent with the semantic area where it is located due to the global patch. However, with our semantic correlation recovery, this dilemma can be greatly improved, and surprising semantic correlations can also be observed among multiple objects with same semantics that are located at a considerable distance from each other.

Figure 3 shows the clustering effect of our model and the noise problem caused by the global patch. We perform denoising operations on this basis, and obtain masks with better qualities, finally finish the predictions of semantic segmentation. In the visualization, the result after denoising is actually the final semantic segmentation results, here we do not mark specific category labels on the images for simplicity.

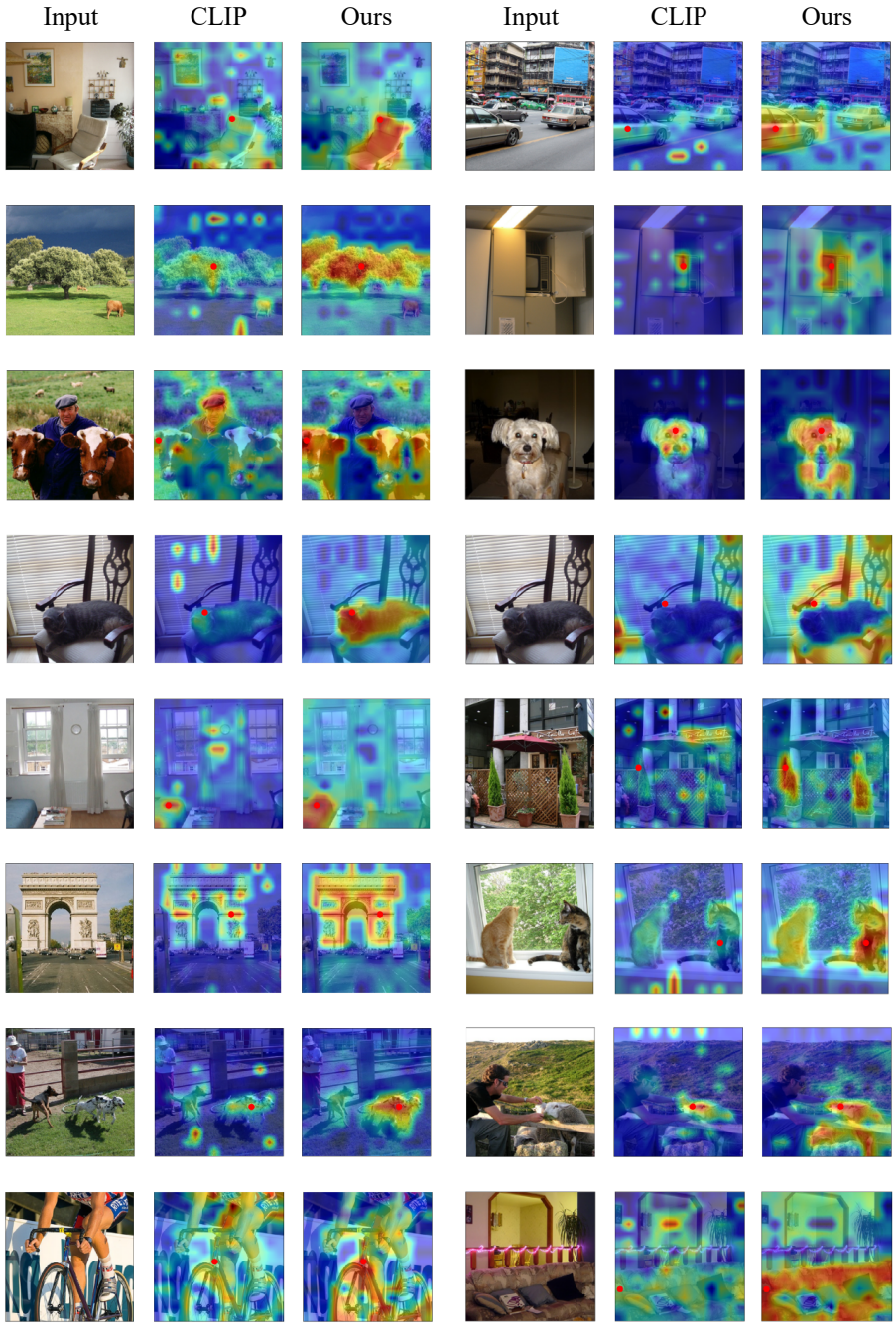

Fig. 2: More comparisons before and after semantic correlation recovery. The red dot indicates the selected patch position.

Fig. 3: Visualization of clustering results and denoising results in our method. The red box represents the noise caused by the existing global patches.

References

1. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
3. Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11165–11174 (2023)
4. Chen, J., Zhu, D., Qian, G., Ghanem, B., Yan, Z., Zhu, C., Xiao, F., Elhoseiny, M., Culatana, S.C.: Exploring open-vocabulary semantic segmentation without human labels. arXiv preprint arXiv:2306.00450 (2023)
5. Chen, Y., Xu, X., Tian, Z., Jia, J.: Homomorphic latent space interpolation for unpaired image-to-image translation. In: CVPR. pp. 2408–2416 (2019)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
7. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
8. Cho, J.H., Mall, U., Bala, K., Hariharan, B.: Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16794–16804 (2021)
9. Cho, S., Shin, H., Hong, S., An, S., Lee, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. arXiv preprint arXiv:2303.11797 (2023)
10. Contributors, P.: Pointcept: A codebase for point cloud perception research. `https://github.com/Pointcept/Pointcept` (2023)
11. Cui, J., Liu, S., Tian, Z., Zhong, Z., Jia, J.: Reslt: Residual learning for long-tailed recognition. TPAMI **45**(3), 3695–3706 (2023)
12. Cui, J., Zhong, Z., Tian, Z., Liu, S., Yu, B., Jia, J.: Generalized parametric contrastive learning. CoRR **abs/2209.12400** (2022)
13. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11583–11592 (2022)
14. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary panoptic segmentation with maskclip. arXiv preprint arXiv:2208.08984 (2022)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015)

17. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European Conference on Computer Vision. pp. 540–557. Springer (2022)
18. Guo, J., Wang, Q., Gao, Y., Jiang, X., Lin, S., Zhang, B.: Mvp-seg: Multi-view prompt learning for open-vocabulary semantic segmentation. In: Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV 2023, Xiamen, China, October 13–15, 2023, Proceedings, Part XII. p. 158–171. Springer-Verlag (2023)
19. Han, C., Zhong, Y., Li, D., Han, K., Ma, L.: Open-vocabulary semantic segmentation with decoupled one-pass network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1086–1096 (2023)
20. Jiang, L., Shi, S., Tian, Z., Lai, X., Liu, S., Fu, C., Jia, J.: Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In: ICCV. pp. 6403–6412 (2021)
21. Jiao, S., Wei, Y., Wang, Y., Zhao, Y., Shi, H.: Learning mask-aware clip representations for zero-shot segmentation. Advances in Neural Information Processing Systems **36** (2024)
22. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
23. Kim, C., Han, W., Ju, D., Hwang, S.J.: Eagle: Eigen aggregation learning for object-centric unsupervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3523–3533 (2024)
24. Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., Jia, J.: Semi-supervised semantic segmentation with directional context-aware consistency. In: CVPR. pp. 1205–1214 (2021)
25. Lan, M., Wang, X., Ke, Y., Xu, J., Feng, L., Zhang, W.: Smooseg: smoothness prior for unsupervised semantic segmentation. Advances in Neural Information Processing Systems **36** (2024)
26. Lee, D., Song, S., Suh, J., Choi, J., Lee, S., Kim, H.J.: Read-only prompt optimization for vision-language few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1401–1411 (2023)
27. Li, Y., Wang, H., Duan, Y., Li, X.: Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv preprint arXiv:2304.05653 (2023)
28. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
29. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15305–15314 (2023)
30. Liu, J., Yang, S., Jia, P., Lu, M., Guo, Y., Xue, W., Zhang, S.: Vida: Homeostatic visual domain adapter for continual test time adaptation. arXiv preprint arXiv:2306.04344 (2023)
31. Liu, Y., Bai, S., Li, G., Wang, Y., Tang, Y.: Open-vocabulary segmentation with semantic-assisted calibration. arXiv preprint arXiv:2312.04089 (2023)
32. Luo, X., Tian, Z., Zhang, T., Yu, B., Tang, Y.Y., Jia, J.: Pfenet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask. TPAMI **46**(2), 1273–1289 (2024)

33. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014)
34. Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19413–19423 (2023)
35. Peng, B., Tian, Z., Wu, X., Wang, C., Liu, S., Su, J., Jia, J.: Hierarchical dense correlation distillation for few-shot segmentation. In: CVPR. pp. 23641–23651 (2023)
36. 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)
37. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18082–18091 (2022)
38. Seong, H.S., Moon, W., Lee, S., Heo, J.P.: Leveraging hidden positives for unsupervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19540–19549 (2023)
39. Shi, H., Hayat, M., Wu, Y., Cai, J.: Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9611–9620 (2022)
40. Shu, Y., Guo, X., Wu, J., Wang, X., Wang, J., Long, M.: Clipood: Generalizing clip to out-of-distributions. arXiv preprint arXiv:2302.00864 (2023)
41. Tang, J., Zheng, G., Shi, C., Yang, S.: Contrastive grouping with transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23570–23580 (2023)
42. Tang, L., Li, K., He, C., Zhang, Y., Li, X.: Consistency regularization for generalizable source-free domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4323–4333 (2023)
43. Tang, L., Li, K., He, C., Zhang, Y., Li, X.: Source-free domain adaptive fundus image segmentation with class-balanced mean teacher. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 684–694. Springer (2023)
44. Tang, L., Tian, Z., Li, K., He, C., Zhou, H., Zhao, H., Li, X., Jia, J.: Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. arXiv preprint arXiv:2407.05342 (2024)
45. Tian, Z., Lai, X., Jiang, L., Liu, S., Shu, M., Zhao, H., Jia, J.: Generalized few-shot semantic segmentation. In: CVPR. pp. 11553–11562 (2022)
46. Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X., Jia, J.: Learning shape-aware embedding for scene text detection. In: CVPR. pp. 4234–4243 (2019)
47. Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J.: Prior guided feature enrichment network for few-shot segmentation. TPAMI **44**(2), 1050–1065 (2022)
48. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. arXiv preprint arXiv:2312.01597 (2023)
49. Wei, Y., Cao, Y., Zhang, Z., Yao, Z., Xie, Z., Hu, H., Guo, B.: icar: Bridging image classification and image-text alignment for visual recognition. arXiv preprint arXiv:2204.10760 (2022)

50. Wu, L., Zhang, W., Jiang, T., Yang, W., Jin, X., Zeng, W.: [cls] token is all you need for zero-shot semantic segmentation. arXiv preprint arXiv:2304.06212 (2023)
51. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: The Twelfth International Conference on Learning Representations (2024)
52. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022)
53. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
54. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
55. Yang, S., Liu, J., Zhang, R., Pan, M., Guo, Z., Li, X., Chen, Z., Gao, P., Guo, Y., Zhang, S.: Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. arXiv preprint arXiv:2312.14074 (2023)
56. Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J.: An improved baseline for reasoning segmentation with large language model. arXiv preprint arXiv:2312.17240 (2023)
57. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1020–1031 (2023)
58. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)
59. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)
60. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
61. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
62. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)