1 Limitations.

The training guidance derived from patch-wise representation still has limitations in capturing intricate pixel-level details, especially along object edges. This issue becomes more pronounced with larger patch sizes, as ViT-S/16 exhibits lower mIoU compared to ViT-S/8 in this regard.

2 Inference

We introduce the overall flow of our model at the inference stage in Fig. 1. During training, we finetune the last block of the Vision Transformer (ViT) and the projection head which outputs the projected vector $\mathbf{z}$. While the projected vector $\mathbf{z}$ is used for the training, we use the feature from the backbone for the inference as did in [1–3].

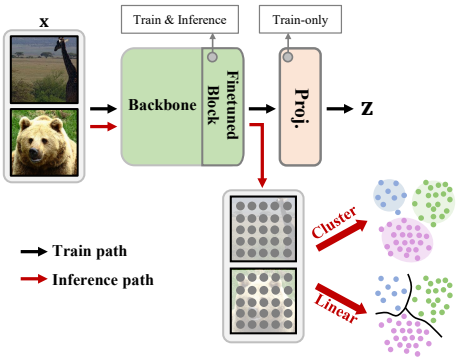

Fig. 1: Illustration of the training and inference process.

3 Implementation Details

In line with existing works [4, 9, 11], we utilize a DINO-pretrained ViT as our backbone network. The embedding dimension of the projection layer (D) is 384 for ViT-small and 768 for ViT-base. We set T to 2, 3, and 1 for the experiments on COCO-stuff, Cityscapes, and Potsdam-3 datasets, respectively. Tab. 1 shows the hyperparameter set that yields the best performance for each dataset and backbone. To reduce the training cost, we use only a random subset of the patch features [7]. Specifically, we use 1/4 of all patch features for the ViT-S/16 backbone and 1/16 for the other backbones. For the experiments on the ImageNet-S dataset, we use the same hyperparameter settings as for the COCO-stuff dataset.

Table 1: Hyperparameter settings for each experiment.

	COCO-stuff		Cityscapes		Potsdam-3
	ViT-S/8	ViT-S/16	ViT-S/8	ViT-B/8	ViT-B/8
Φ	0.55	0.55	0.6	0.6	0.55
Ψ	0.2	0.15	0.2	0.2	0.15
σ_{pos}	3	3	3	3	5
σ_{amb}	3	4	3	2	3

4 Additional Study

4.1 Study on Effects to Class Frequency

In dense prediction tasks, the variation in the frequency between different semantics is a very natural phenomenon. However, this typically leads to the problem of long-tailed data distribution which triggers the large performance gap between classes of high and low frequencies [10]. To further analyze the strength of our proposed method, we compare with HP [9] on the effects from the perspective of the class frequency. Specifically, we divide the classes in the COCO-stuff dataset into three groups, i.e., Few, Medium, and Many, according to the data frequency and measure the performances for each group. Results are reported in Tab. 2. As shown, we find that our method is particularly notable in learning classes with fewer samples compared to the baseline. For such a result, we attribute a reason that the number and the precision of the gathered positives are similar between samples as shown in Fig. 1 of the main paper.

Table 2: Experimental results considering long-tailed distribution.

Method	Few	Medium	Many	All
HP	28.9	47.1	75.6	42.0
Ours	32.9	50.9	76.2	45.6

4.2 Different Pretrained Backbones.

The results in Tab. 3 demonstrate consistent performance improvements with various backbones pretrained in a self-supervised manner (*e.g.*, iBoT [13], SelfPatch [12]). We observed that backbones trained with inter-image relationships consistently enhance performance. However, models such as MAE [5] struggle to preserve globally shared semantics in each patch feature across all images (*e.g.*, 4.3% of U.mIoU for MAE alone) due to their lack of inter-image relationship modeling [6].

Table 3: Experimental results with various pretrained backbones.†: reproduced.

Method	Backbone	Unsupervised Acc.	Unsupervised mIoU
iBoT [13]	ViT-S/16	39.2	11.8
+ HP† [9]	ViT-S/16	53.2	23.0
+ PPAP (Ours)	ViT-S/16	**62.4**	**26.0**
iBoT [13]	ViT-B/16	35.7	15.0
+ HP† [9]	ViT-B/16	51.1	22.4
+ PPAP (Ours)	ViT-B/16	**63.4**	**27.6**
SelfPatch [12]	ViT-S/16	35.1	12.3
+ STEGO [4]	ViT-S/16	52.4	22.2
+ HP [9]	ViT-S/16	56.1	23.2
+ PPAP (Ours)	ViT-S/16	**57.8**	**23.7**

4.3 Contribution of CRF

Conditional Random Field (CRF) utilizes pixel position and RGB color information to smooth the predicted label of each pixel across its neighboring pixels, thereby effectively enhancing the performance of semantic segmentation [8]. Following the previous works [4,9], we incorporate CRF as a post-processing step in our method. Tab. 4 presents a performance comparison between HP [9] and our method, both with and without the application of CRF. While the use of CRF leads to performance boosts in all experiments, we highlight the superiority of our PPAP in that it outperforms HP with CRF even without CRF.

Table 4: Experimental results on COCO-stuff dataset with and without CRF.

Backbone	Method	CRF	Unsupervised Acc.	Unsupervised mIoU	Linear Acc.	Linear mIoU
ViT-S/8	HP [9]	-	56.2	23.9	73.7	40.4
		✓	57.2	24.6	75.6	42.7
	Ours	-	57.4	26.6	76.0	45.4
		✓	59.0	27.2	76.9	46.3
ViT-S/16	HP [9]	-	52.7	23.2	71.9	37.1
		✓	54.5	24.3	74.1	39.1
	Ours	-	59.9	25.0	74.3	42.1
		✓	62.9	26.5	76.0	43.3

4.4 Qualitative Results without CRF

We provide visualizations comparing the predictions of our proposed PPAP with those of existing methods, *i.e.*, STEGO [4], and HP [9], without CRF. The results without CRF are shown in Fig. 2 and 3. As can be noticed, we point out that existing methods tend to be prone to noise, whereas our method demonstrates robustness against pixel-wise noises even without applying the CRF.

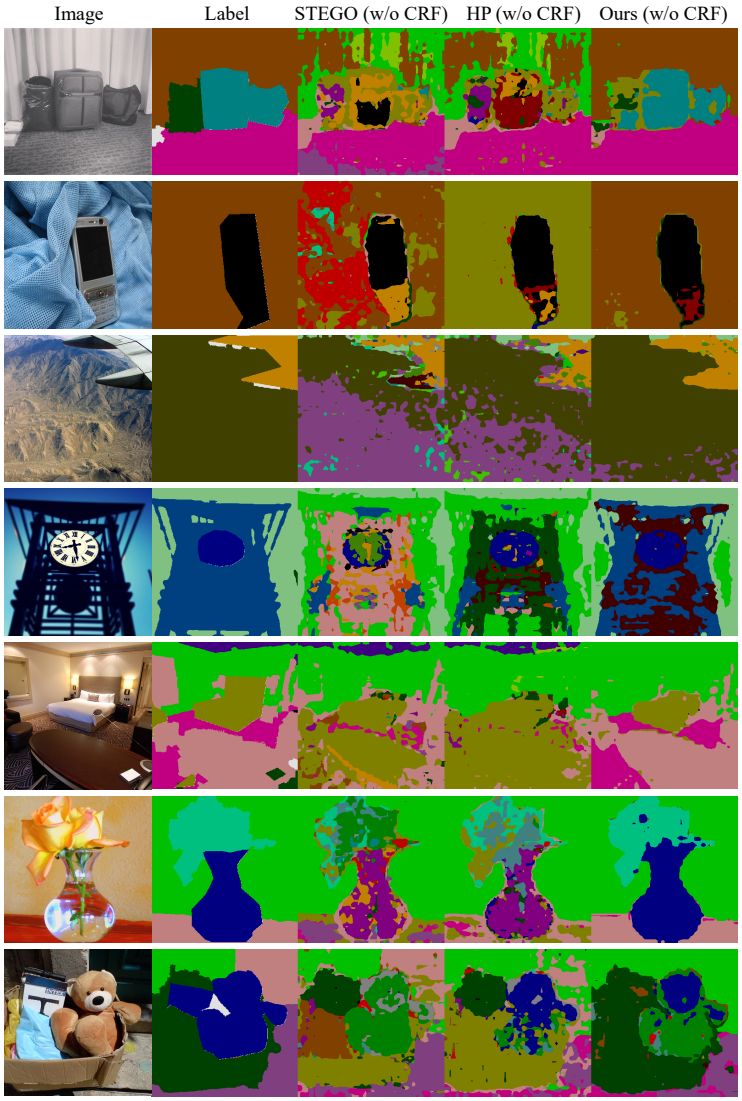

Fig. 2: Qualitative results without CRF.

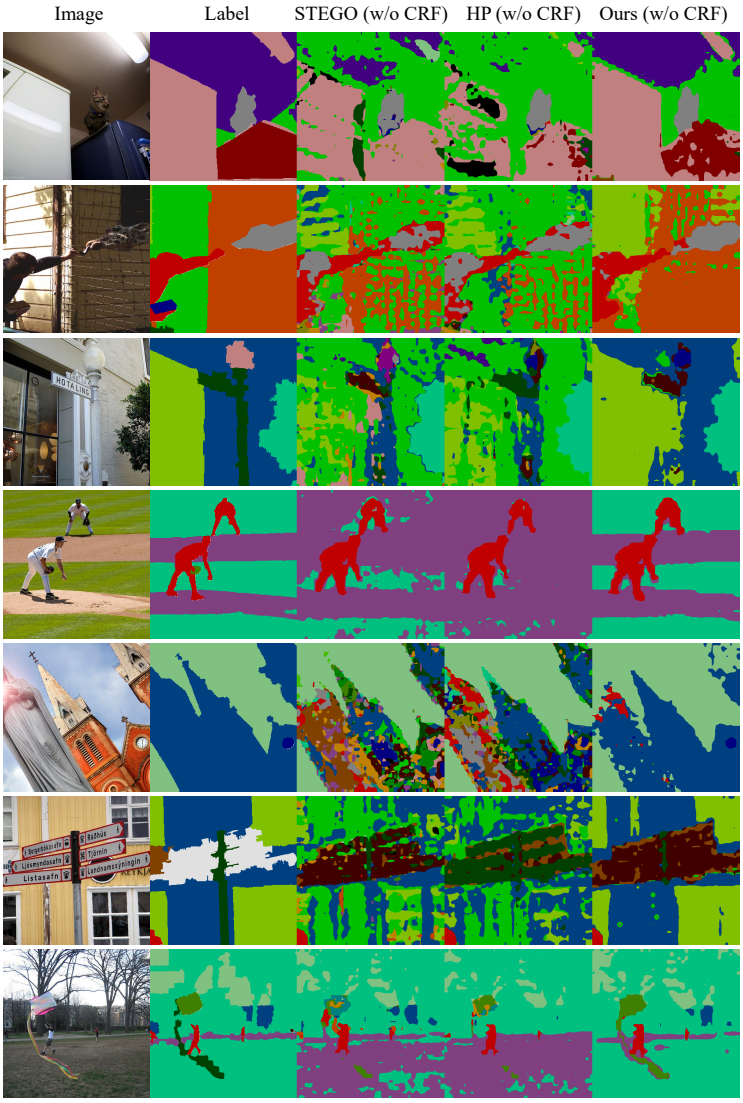

Fig. 3: Qualitative results without CRF.

References

1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)
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. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=SaKO6z6Hl0c`
5. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
6. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
7. Kim, J., Lee, B.K., Ro, Y.M.: Causal unsupervised semantic segmentation. arXiv preprint arXiv:2310.07379 (2023)
8. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. Advances in neural information processing systems **24** (2011)
9. 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)
10. Wang, Y., Fei, J., Wang, H., Li, W., Bao, T., Wu, L., Zhao, R., Shen, Y.: Balancing logit variation for long-tailed semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19561–19573 (2023)
11. Yin, Z., Wang, P., Wang, F., Xu, X., Zhang, H., Li, H., Jin, R.: Transfgu: a top-down approach to fine-grained unsupervised semantic segmentation. In: European Conference on Computer Vision. pp. 73–89. Springer (2022)
12. Yun, S., Lee, H., Kim, J., Shin, J.: Patch-level representation learning for self-supervised vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8354–8363 (2022)
13. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)