Appendix

Additional ablations. We have observed that training CP^2 without copy-paste and dense contrastive loss yields poor downstream performance. To further ablate this effect, we pretrain CP^2 models with copy-pasted images but using only one loss term, *i.e.* the dense contrastive loss $\mathcal{L}_{dense}$ or the instance contrastive loss $\mathcal{L}_{ins}$. This ablation study further dissects the components of CP^2 and gives better comprehension of their effects. For better efficiency, we conduct this ablation on top of the MoCo v2 800-epoch ResNet-50 and apply CP^2 Quick Tuning for 20 epochs. As listed in Table 6, compared with the MoCo v2 baseline, applying Quick Tuning protocol in the MoCo manner (no copy-paste) slightly improves the downstream performance by 0.4% mIoU, while employing the entire CP^2 design yields 1.4% mIoU improvements.

Table 6: **Ablation results** of CP^2. The results (mIoU on PASCAL VOC) are based on ResNet50-ASPP models, where the base backbone is loaded from the MoCo v2 pretrained ResNet50 for 800 epochs.

mode	QT	copy-paste	dense loss	instance loss	mIoU
baseline		-	-	-	77.2
no copy-paste	✓			✓	77.6
instance loss only	✓	✓		✓	78.0
dense loss only	✓	✓	✓		75.6
entire CP^2	✓	✓	✓	✓	78.6

We also note that despite omitting the dense contrastive loss of CP^2 (instance loss only) leads to 0.6% mIoU degradation (compared with entire CP^2), it outperforms Quick Tuning with no copy-pasted images. The only difference between the setup of *no copy-paste* and *instance loss only* is that the former uses original augmented views and apply **global average pooling** to the dense features for instance discrimination, while the later uses copy-pasted views and **masked pooling**. Therefore, this result indicates that training with copy-pasted images provides the models with better robustness to background noise, which benefits the downstream performance of semantic segmentation. Moreover, omitting the instance contrastive loss of CP^2 yields even worse downstream performance (dense loss only) than the MoCo v2 baseline, which indicates that the capability of instance discrimination is one of the models' key components for semantic segmentation.

Deviations of main results. To demonstrate the robustness of CP^2, here we in addition report the standard deviations of our main experimental results. We run three trials of CP^2 Quick Tuning and downstream finetuning and list the mean±std results in Table 7.

Table 7: **Deviations of CP^2 results**. The results are based on ResNet50-ASPP models, where the base backbone is loaded from the MoCo v2 pretrained ResNet50 for 800 epochs. We run three trials for mean±std results.

dataset	PASCAL	Cityscapes	ADE20k
mIoU	78.6±0.2	77.4±0.2	41.3±0.1