Supplementary Material

A. More Implementation Details

The experiments in Tab. 2 of the main paper are conducted on a single NVIDIA Tesla A100 while the remaining experiments are conducted on 2 NVIDIA Tesla A5000s with the Distributed DataParallel (DDP) mode. Meanwhile, we apply random cropping, random horizontal flipping, and random scaling with the range of [0.5, 2.0] for both VOC 2012 and CityScapes, following previous works [2]. Cutmix [1] is used as a strong data augmentation in our method.

The hyper-parameter λ_c in Eq. 6 of the main paper is a time-variant scaling parameter $\lambda_c(t)$, which is formulated as:

$$\lambda_c(t) = \lambda_{c0} \cdot exp\left(\gamma \cdot (\frac{t}{T_{total}})^2\right), \tag{1}$$

where λ_{c0} denotes the initial scaling parameter, γ denotes a weight decay coefficient, t denotes the current t^{th} epoch and T_{total} denotes the total epochs.

All of the hyper-parameters are shown in Tab. 1.

Table 1. All of the hyper-parameters in CSS

Symbol	Description	Default
α_t	update speed of teacher model	0.99
α_p	update speed of prototypes	0.99
δ_u	threshold for sampling strategy in $\mathcal{L}_u$	0.97
λ_{c0}	initial scaling parameter of $\lambda_c(t)$	1.0
γ	weight decent coefficient in $\lambda_c(t)$	-5.0
τ	temperature for contrastive loss $\mathcal{L}_c$	0.5
δ_w	threshold for sampling strategy in $\mathcal{L}_c$	0.7
δ_s	threshold for sampling strategy in $\mathcal{L}_c$	0.8
-	warm-up epochs	20

B. More Quantitative results

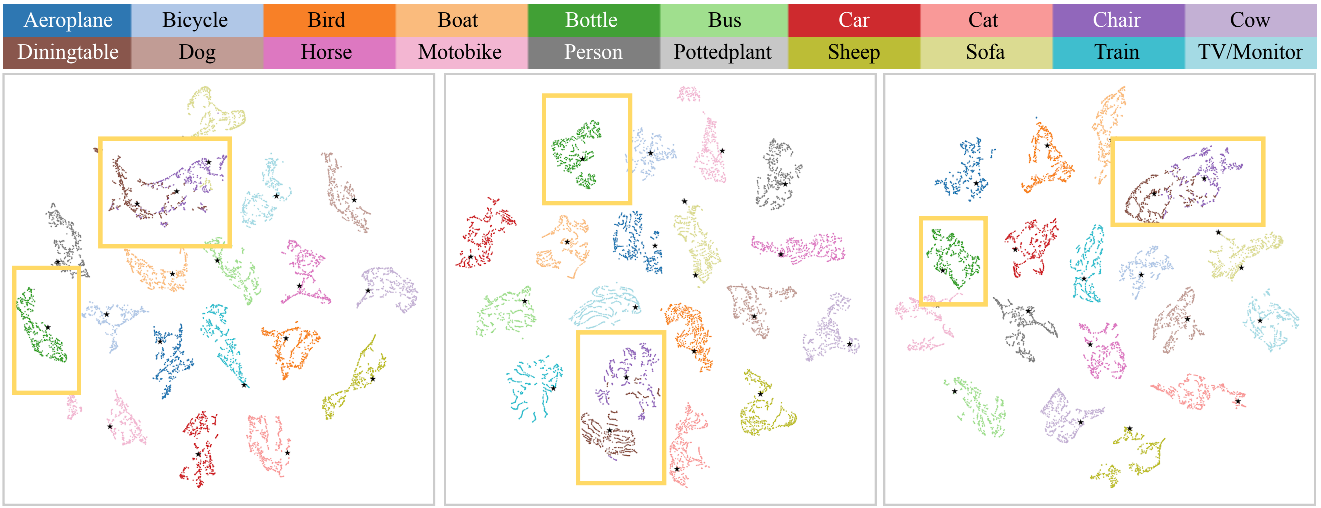
We report the IoU of three methods (Baseline, CSS (mix), and CSS (cross)) on PASCAL VOC 2012 with 92 labels in Tab. 2. The results are produced on one data split and thus differ from Tab. 1 in the main paper. Even though our method degrades the performance of some classes (*e.g.*, `cat`, `dog`, and `train`), the IoU of those under-performing classes (*e.g.*, `potted plant`, and `bottle`) in the baseline is dramatically boosted. We mainly attribute it to the knowledge exchange between the logit and representation spaces. Pseudo-labels from different spaces help the model learn the concentrations of different spaces and obtain a more balanced performance in each class.

Tab. 3 shows the IoU of our method on Cityscapes dataset with different label rates.

C. More Qualitative Results

Fig. 1 shows the t-SNE [3] visualization of baseline and CSS. Thanks to the dual-space collaborative supervision, representations of the same class in our CSS are more compact than that in the baseline.

Fig. 2 shows the qualitative results of different methods on Cityscapes with 186 labeled images. Baseline means the conventional contrastive-based method. Compared with our baselines, benefiting from the supervision of two spaces and different indicators in different spaces, our method performs better.



(a) Baseline

(b) CSS (mix)

(c) CSS (cross)

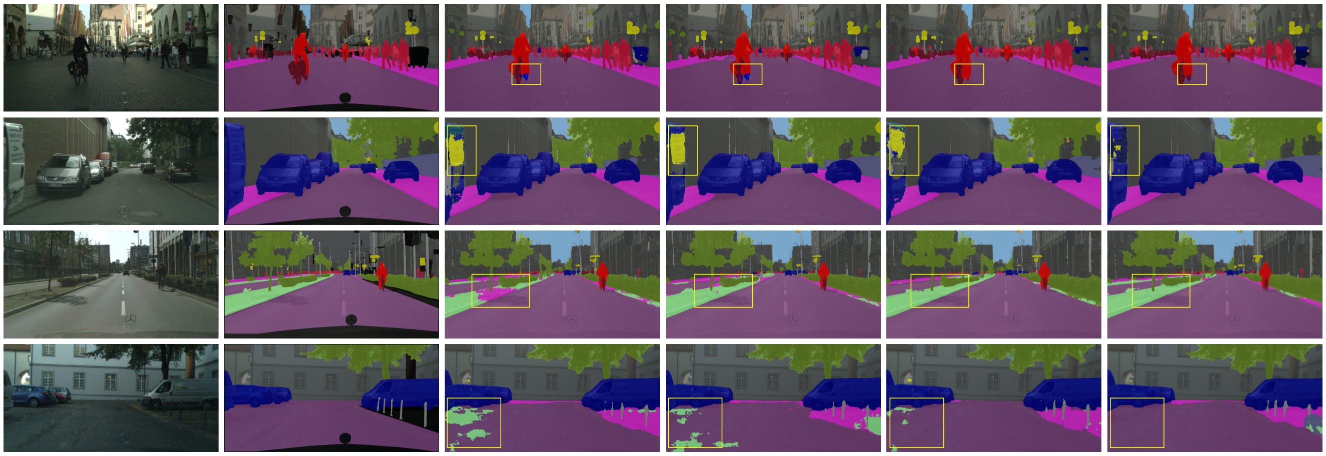
Figure 1. T-SNE visualization of baseline and our CSS. Black stars mean prototypes. Yellow boxes highlight the main difference.

Table 2. The IoU of each class in PASCAL VOC 2012 dataset with 92 labeled images.

method	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
Baseline	91.69	75.08	50.24	80.24	61.20	47.34	86.46	80.41	87.14	17.55	77.97
cross	$91.68_{\downarrow 0.01}$	$74.58_{\downarrow 0.49}$	$52.73_{\uparrow 2.50}$	$83.53_{\uparrow 3.29}$	$68.72_{\uparrow 7.51}$	$56.68_{\uparrow 9.35}$	$86.47_{\uparrow 0.01}$	$80.73_{\uparrow 0.32}$	$86.18_{\downarrow 0.96}$	$17.06_{\downarrow 0.49}$	$82.32_{\uparrow 4,34}$
mix	$92.71_{\uparrow 1.02}$	$84.01_{\uparrow 8.94}$	$58.25_{\uparrow 8.02}$	$70.31_{\downarrow 7.22}$	$66.82_{\uparrow 5.61}$	$65.51_{\uparrow 18.17}$	$83.28_{\downarrow 3.18}$	$84.17_{\uparrow 3.76}$	$81.17_{\downarrow 5.97}$	$26.57_{\uparrow 9.02}$	$73.92_{\downarrow 4.05}$
method	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tv/monitor	mIoU
Baseline	46.72	82.72	79.49	66.45	81.53	38.43	74.11	33.38	80.88	59.11	66.58
cross	$43.55_{\downarrow 3.17}$	$79.03_{\downarrow 3.69}$	$77.54_{\downarrow 1.96}$	$71.27_{\uparrow 4.82}$	$80.70_{\downarrow 0.83}$	$42.88_{\uparrow 4.45}$	$81.26_{\uparrow 7.15}$	$40.74_{\uparrow 7.36}$	$79.38_{\downarrow 1.50}$	$55.41_{\downarrow 3.70}$	$68.21_{\uparrow 1.63}$
mix	$55.58_{\uparrow 8.86}$	$73.48_{\downarrow 9.24}$	$72.01_{\downarrow 7.48}$	$76.91_{\uparrow 10.47}$	$80.86_{\downarrow 0.67}$	$47.19_{\uparrow 8.76}$	$79.00_{\uparrow 4.89}$	$35.08_{\uparrow 1.69}$	$78.98_{\downarrow 1.90}$	$65.08_{\uparrow 5.97}$	$69.22_{\uparrow 2.64}$

Table 3. The IoU of each class in Cityscapes dataset with four label rates. `vege.` denotes class `vegetation`, `T. light` denotes class `traffic light`, and `T. sign` denotes class `traffic sign`.

label	road	sidewalk	building	fence	pole	vege.	terrain	sky	person	car
186	98.02	80.29	90.06	56.16	52.73	60.14	67.11	74.67	91.20	64.52
372	98.07	83.66	93.93	58.33	53.88	67.85	69.56	76.28	91.31	65.65
744	98.10	83.97	93.30	60.31	55.75	67.52	70.60	76.85	92.08	67.46
1488	98.69	86.69	93.63	60.14	59.33	68.15	70.71	77.86	93.02	68.54
label	wall	T. light	T. sign	rider	truck	bus	train	motor.	bicycle	mIoU
186	93.07	79.64	61.96	93.30	58.88	81.23	69.37	60.05	73.91	74.02
372	93.87	80.38	65.04	93.54	66.05	87.91	72.79	66.80	76.98	76.94
744	93.71	81.28	66.23	93.59	70.99	87.93	75.46	68.41	77.45	77.95
1488	93.99	83.79	70.30	93.81	73.87	89.94	77.56	74.13	78.51	79.63



(a) Image (b) GT (c) CutMix (d) Baseline (e) CSS (cross) (f) CSS (mix)

Figure 2. Visualization on Cityscapes with 186 labeled images. Yellow boxes highlight the main differences.

References

[1] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *In 31th British Machine Vision Conference*, 2020.

[2] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew Davison. Bootstrapping semantic segmentation with regional contrast. In *International Conference on Learning Representations*, 2022.

[3] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.