SUPPLEMENTARY MATERIAL

This document provides additional details to support the main paper, including dataset statistics, full hyperparameter settings, formal proof, extended training protocols, and additional ablation studies.

A DATASET SPLITS

Table 1 summarizes the datasets used in our experiments. We use a 10% labelled split of Cityscapes' 2 975 training images (298 labeled / 2 677 unlabeled) and a stratified 20% split of ADE20K's 20 210 training images (1 000 labeled / 2 537 unlabeled). Standard validation sets are retained (500 images for Cityscapes, 2 000 for ADE20K). Exact image-ID lists will be released with our code.

Table 1: Semi-supervised splits used in our experiments.

Dataset	# Classes	Labeled / Unlabeled	Validation
Cityscapes	8	298 / 2 677	500
ADE20K	100	1 000 / 2 537	2 000

B HYPERPARAMETERS

Key teacher and student hyperparameters are summarized in Table 2. Results are averages over three independent runs with different random seeds.

Table 2: Hyperparameter Settings

Parameter	**Teacher**	**Student**
Learning rate	5.0×10^{-5}	Encoder: 5.0×10^{-6}; Decoder: 5.0×10^{-5}
Scheduler	Multi-step (milestones at 0.9, 0.95)	PolyLR (power 0.9)
Batch size	4	8
Weight decay	0.01	0.05
Contrastive loss weight	0.2	0.2
Pseudo-label threshold	0.3	0.3
Dropout rate	—	0.1
Gradient clipping	—	ℓ_2 norm 0.1
Optimizer	AdamW (β1=0.9, β2=0.999)	
Augmentations	Weak: flip, resize; Strong: random resized crop, jitter, grayscale, blur,	
Loss weights (mask / class)	5 / 2	

C PROOF SKETCH OF PROPOSITION 3.1

Proof Sketch. Let z_a, z^+ and $\{z_r^-\}_{r=1}^R$ be the unit norm embeddings of an anchor pixel, its positive, and R negatives. Define

$$s^+ = \langle z_a,\, z^+ \rangle, \qquad s_r^- = \langle z_a,\, z_r^- \rangle,$$

and the pixel-wise contrastive loss

$$\ell(z_a) = -\log \frac{\exp(s^+)}{\exp(s^+) + \sum_{r=1}^R \exp(s_r^-)}.$$

Let

$$Z = \exp(s^+) + \sum_{r=1}^R \exp(s_r^-), \qquad \alpha_r = \frac{\exp(s_r^-)}{Z}.$$

A straightforward gradient computation gives

$$\nabla_{z_a} \ell = \sum_{r=1}^R \alpha_r \,(z_r^- - z^+).$$

Applying one gradient descent step with step size λ_{pxl}:

$$z_a' = z_a - \lambda_{\text{pxl}} \nabla_{z_a} \ell \;=\; z_a + \lambda_{\text{pxl}} \sum_{r=1}^{R} \alpha_r \, (z^+ - z_r^-).$$

For a randomly chosen negative z^-,

$$\Delta s^+ = \langle z_a' - z_a,\, z^+ \rangle = \lambda_{\text{pxl}} \sum_{r=1}^{R} \alpha_r \big(1 - \langle z_r^-,\, z^+ \rangle\big),$$

$$\Delta s^- = \langle z_a' - z_a,\, z^- \rangle = \lambda_{\text{pxl}} \sum_{r=1}^{R} \alpha_r \big(\langle z^+,\, z^- \rangle - \langle z_r^-,\, z^- \rangle\big).$$

By Assumption 3.1, each negative embedding z_r^- is inter-instance with probability p, in which case $\langle z_r^-,\, z^+ \rangle \approx 0$, and intra-instance with probability $1 - p$, in which case $\langle z_r^-,\, z^+ \rangle \approx 1$. Hence

$$\mathbb{E}\big[1 - \langle z_r^-,\, z^+ \rangle\big] = p \cdot 1 + (1 - p) \cdot 0 = p,$$

and since $\sum_{r=1}^{R} \alpha_r = 1$, it follows that

$$\mathbb{E}[\Delta s^+] = \lambda_{\text{pxl}} \sum_{r=1}^{R} \alpha_r \, \mathbb{E}\big[1 - \langle z_r^-,\, z^+ \rangle\big] = p \, \lambda_{\text{pxl}}.$$

Meanwhile, every term in Δs^- involves an inter-instance inner product, either $\langle z^+, z^- \rangle$ or $\langle z_r^-, z^- \rangle$ each of which vanishes in expectation, so $\mathbb{E}[\Delta s^-] \approx 0$. Therefore

$$\mathbb{E}[\Delta s^+ - \Delta s^-] = p \, \lambda_{\text{pxl}} - 0 = \Theta\big(p \, \lambda_{\text{pxl}}\big) = \varepsilon > 0,$$

i.e. one update on $\mathcal{L}_{\text{pxl}}$ increases the expected inter-instance margin by ε. □

Remark C.1 (Why $\langle z^+, z^- \rangle \approx 0$ holds)**.** *Under the InfoNCE objective (§3.2), the normalized weights for negative pairs, $\alpha_r = \frac{e^{s_r^-}}{e^{s^+} + \sum_r e^{s_r^-}}$, vanish at convergence, i.e. $\alpha_r \approx 0$. Moreover, in high dimensional embeddings, random unit vectors have inner products concentrating near zero, and contrastive training further pushes these negative similarities into a tight, small magnitude distribution Chen et al. (2020). Thus it is reasonable to approximate $\langle z^+, z^- \rangle \approx 0$ up to $O(1/\sqrt{D})$ fluctuations.*

D More Training Details

All teacher models are fine-tuned using 1k iterations on labeled set, followed by 5k iterations in a self-training stage with pseudo-labels. For student models, training on the Cityscapes dataset spans 90k iterations, consistent with prior works, while the mini-ADE20k dataset is trained for 80k iterations. Finally, both datasets undergo an additional supervised fine-tuning phase for 2k iterations.

E Additional Ablation Studies

E.1 Ablation: Teacher Adaptation Variants

Different teacher adaptation strategies impact both teacher and student performance. Specifically, we compare fine-tuning only, self-training, and self-training combined with our proposed contrastive loss.

E.2 Loss Variant: InfoNCE vs. Margin Hinge

Replacing our asymmetric InfoNCE (§3.2) with an margin-based hinge loss yields identical maskAP (32.2%) and +0.6 maskAP$_{50}$, at the cost of 1.6× longer training. This evaluates whether enforcing a fixed positive–negative margin can match or improve upon the performance of InfoNCE.

Table 3: Teacher Adaptation Ablation. Teacher/student AP for different adaptation strategies.

Adaptation Variant	Teacher AP	Student AP	Δ vs. SOTA
Fine-tuning only	28.7	32.0	+1.2
Self-training	29.7	32.2	+1.5
Self-training + Contrastive	**30.5**	**33.9**	**+3.1**

Table 4: Loss Variant Ablation. Default InfoNCE vs. margin-based hinge (margin = 0.2).

Loss Variant	maskAP (%)	maskAP$_{50}$ (%)
Asymmetric InfoNCE (§3.2)	32.2	56.5
Margin hinge (m = 0.1)	32.2	57.1

E.3 Ablation: Debias Score Formulation

We evaluate three instantiations of the debias score function s^{deb} (§3.2):

- **Original s^{deb}:** fusion of mask and class confidences (ours).
- $\left(s^{deb}\right)^2$: square each score to amplify the negatives with high confidence.
- $\sqrt{s^{deb}}$: take the square root of each score to temper the bias.

Table 5: Debias Score Formulation Ablation.

Score Variant	maskAP	maskAP$_{50}$
Original	32.2	56.5
Squared	32.0	56.3
Square-root	31.9	56.2

Table 6: Teacher Choice Ablation.

Model	AP	maskAP$_{50}$
Teacher T1 (0-shot)	22.0	42.3
Teacher T2 (adapted)	30.5	56.6
Student under T1	23.8	42.9
Student under T2	**32.2**	**56.5**

E.4 Ablation: Negative Sampling Scope

We evaluate two negative sampling scopes: (i) sampling only within the current mini batch vs. (ii) sampling from a small memory bank of past pixel embeddings (size 10k). Sampling from a memory bank of 10 k embeddings yields a modest performance gain (+0.5 maskAP, +0.8 maskAP$_{50}$) compared to in-batch sampling. However, incurs approximately 2.2× longer training time due to the overhead of maintaining and querying the memory bank.

Table 7: Sampling Scope Ablation. Mini batch only vs. memory bank negatives.

Scope	maskAP (%)	maskAP$_{50}$ (%)
Mini-batch only	32.2	56.5
Memory bank (10k embeddings)	32.7	57.3

E.5 Teacher Choice: Original vs. Adapted

We compare distilling the student from the original VFM teacher (T1, zero-shot) versus our adapted teacher (T2). As shown in Table 6, using the adapted teacher provides a much stronger signal, yielding a +8.4 AP improvement over the student distilled under T1.

E.6 Extended Backbone Comparison

We compare CAST distilled with a DINOv2-S student against Guided Distillation baselines trained with different teacher backbones, including ResNet-50, DINOv2-B, and DINOv2-L.

Table 8: Extended Backbone Comparison. CAST vs. Guided Distillation

Label Fraction	CAST (DINOv2-S)	Guided Dist. (ResNet-50)	Guided Dist. (DINOv2-B)	Guided Dist. (DINOv2-L)
5%	**30.7**	23.9	25.1	28.8
10%	**33.9**	30.8	27.0	33.0
30%	**40.4**	35.6	35.4	39.1

F Use of LLM Statement

We leverage ChatGPT to polish the paper presentation at the sentence level. Specifically, we provided the LLM some of the draft sentences, and asked the LLM if there is a better version of the given sentence

G Additional Qualitative Results

Figure 1 presents additional qualitative examples. The first and third columns show teacher predictions, while the second and fourth columns show the corresponding student predictions.

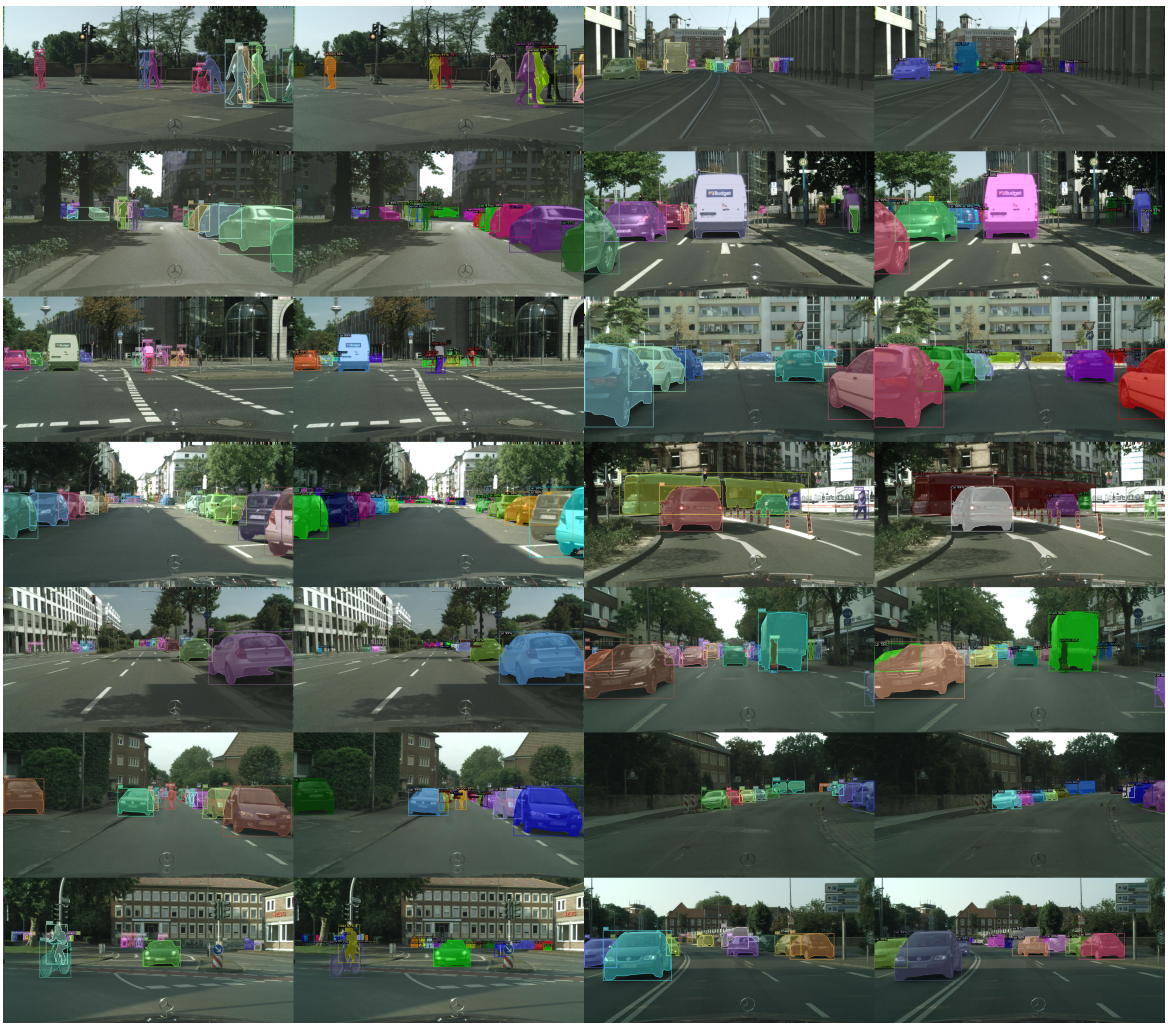

Figure 1: Additional qualitative results on the Cityscapes dataset.