



NeurIPS 2026

D^3S^2

Diffusion-Guided Dataset Distillation for Semantic Segmentation

Wenjie Zheng, Haoji Hu*, Jiali Lu, Xingze Zou, Jing Wang

Zhejiang University

wenjie_zheng@zju.edu.cn

haoji_hu@zju.edu.cn

Dataset distillation for semantic segmentation

A compact training set must preserve **class coverage** and the **spatial structure** of dense labels.

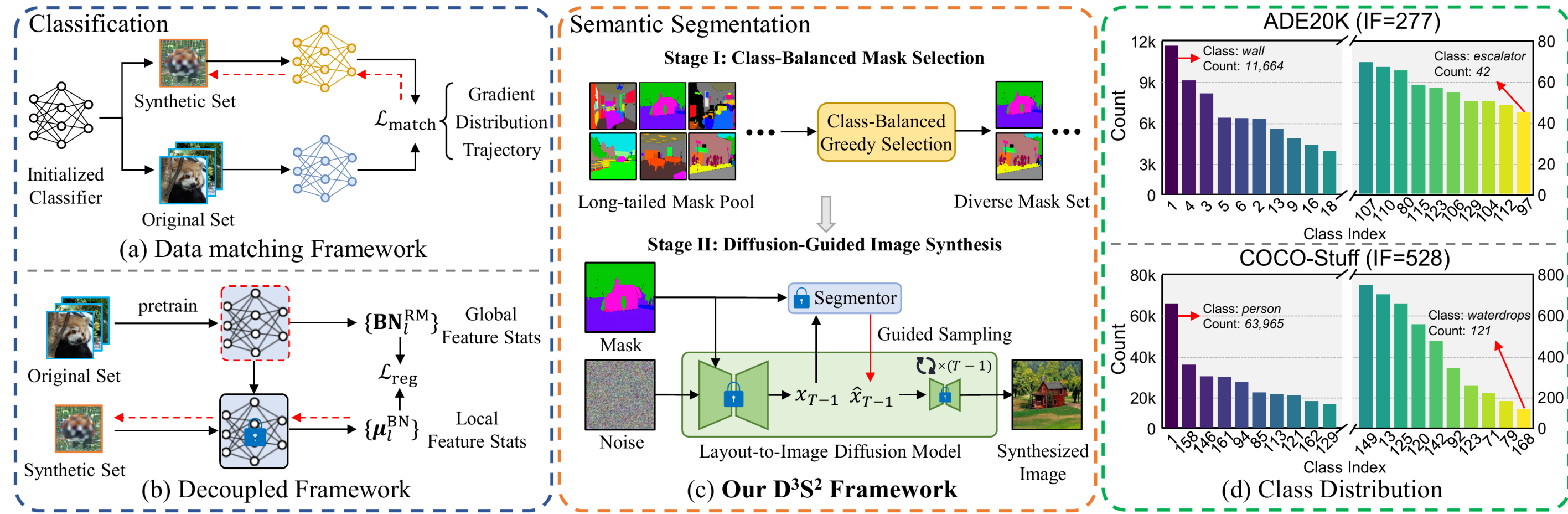


Figure 1. Classification distillation paradigms, D³S², and long-tailed segmentation data

Rare classes

Small budgets can omit semantic categories that appear infrequently.

Spatial alignment

Every synthesized region must agree with its dense segmentation label.

Computation

High-resolution data makes pixel optimization expensive.

D³S² framework

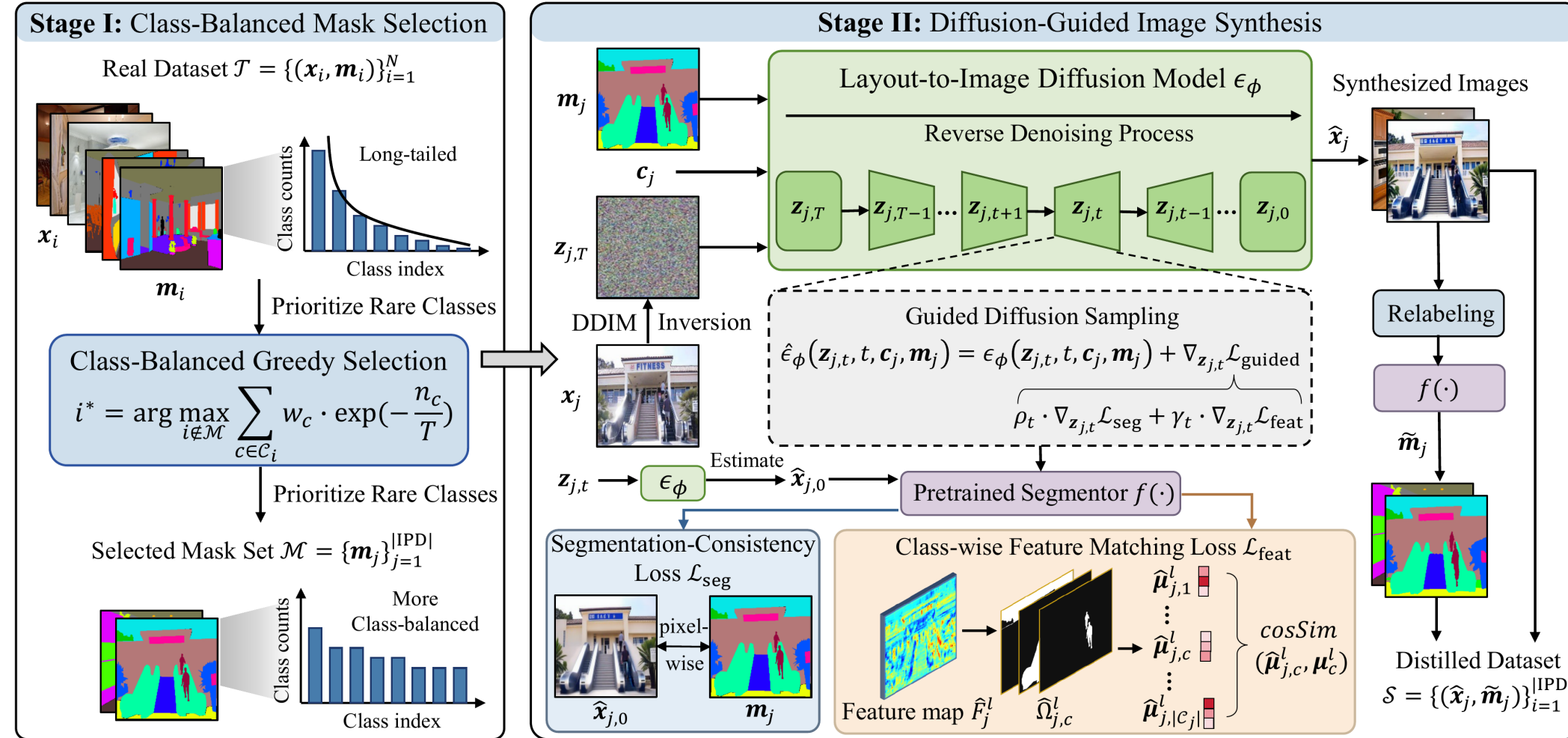


Figure 2. Class-balanced mask selection and diffusion-guided image synthesis

The selected masks define the layouts. Segmentation and class-wise feature losses improve the **training utility** of the synthesized images.

Stage I

Select masks that jointly cover rare and common semantic classes.

Stage II

Refine paired real images through DDIM inversion, guided denoising and relabeling.

Stage I: class-balanced mask selection

$$i^* = \arg \max_{i \notin \mathcal{M}} \sum_{c \in \mathcal{C}_i} w_c \cdot \exp \left(-\frac{n_c}{T} \right)$$

Greedy selection favors masks that improve coverage of underrepresented classes.

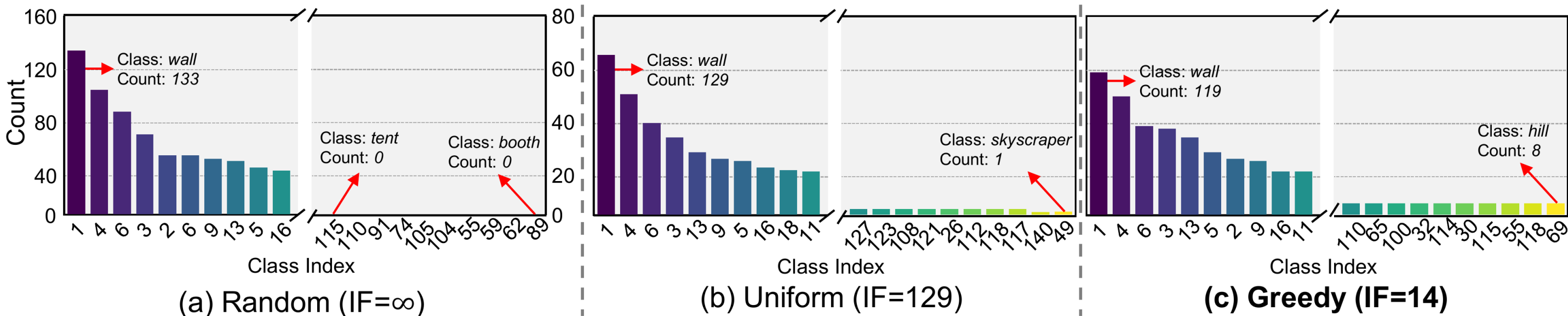
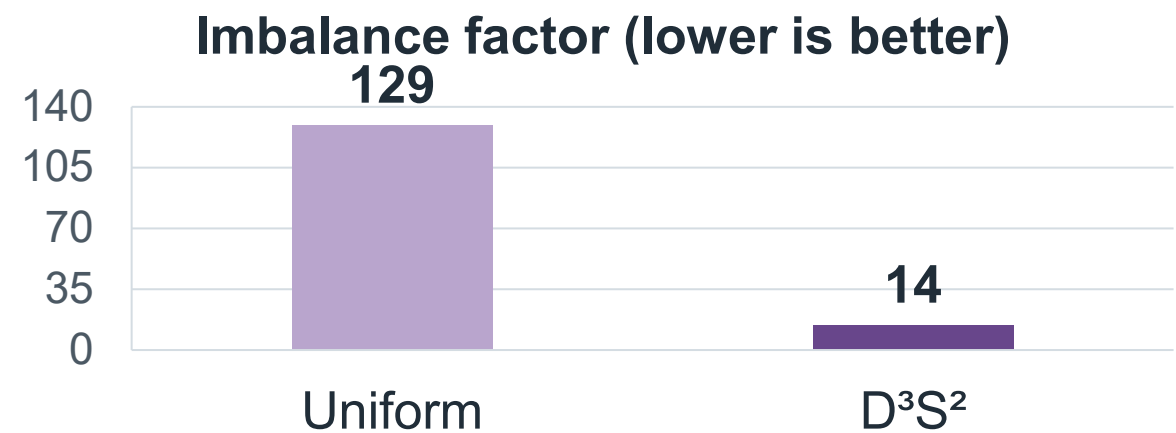


Figure 3. ADE20K class distributions at a 1% data budget



8+

image occurrences for every semantic class

Inverse-frequency weights emphasize rare classes. The coverage term reduces priority once a class already has support.

Stage II: guidance during diffusion sampling

Pixel-level consistency

The segmentation loss compares the segmentor prediction with the selected mask at every denoising step.

$$\mathcal{L}_{\text{seg}}(f(\hat{x}_{j,0}), m_j)$$

This objective preserves object semantics and their spatial boundaries.

Class-wise feature matching

Masked class means match dataset-level feature statistics across network stages. Cosine similarity measures their agreement.

$$\mathcal{L}_{\text{feat}} = \sum_{c \in \mathcal{C}_j} \sum_l (1 - \text{cosSim}(\hat{\mu}_{j,c}^l, \mu_c^l))$$

This objective strengthens class-specific appearance within each semantic region.

Sampling and supervision

DDIM inversion provides the starting latent from the paired real image. Both losses guide 50 denoising steps. **Relabeling** refines the final masks.

Efficiency: about 50 s per image on RTX 6000 Ada (48 GB), excluding pretrained-model training.

ADE20K: gains across segmentation backbones

1% data budget. Mean $\pm$ standard deviation, mIoU (%).

Mask2Former					SegFormer	
Method	ResNet-50	Swin-T	Swin-S	Swin-B	MiT-B2	MiT-B4
Random	12.99 $\pm$ 0.4	13.32 $\pm$ 0.3	15.65 $\pm$ 0.5	18.09 $\pm$ 0.3	16.86 $\pm$ 0.2	19.74 $\pm$ 0.4
K-Center	9.88 $\pm$ 0.4	11.14 $\pm$ 0.3	13.76 $\pm$ 0.4	15.09 $\pm$ 0.2	14.90 $\pm$ 0.3	18.95 $\pm$ 0.3
Herding	13.46 $\pm$ 0.4	14.43 $\pm$ 0.5	16.32 $\pm$ 0.3	17.94 $\pm$ 0.3	16.78 $\pm$ 0.4	19.88 $\pm$ 0.2
Uniform	17.20 $\pm$ 0.3	17.94 $\pm$ 0.2	21.42 $\pm$ 0.3	22.70 $\pm$ 0.4	20.36 $\pm$ 0.3	23.48 $\pm$ 0.4
D³S²	20.96$\pm$0.3	21.69$\pm$0.3	24.99$\pm$0.5	27.48$\pm$0.5	23.82$\pm$0.4	28.17$\pm$0.5

+9.34

points over Random
with Mask2Former (Swin-S)

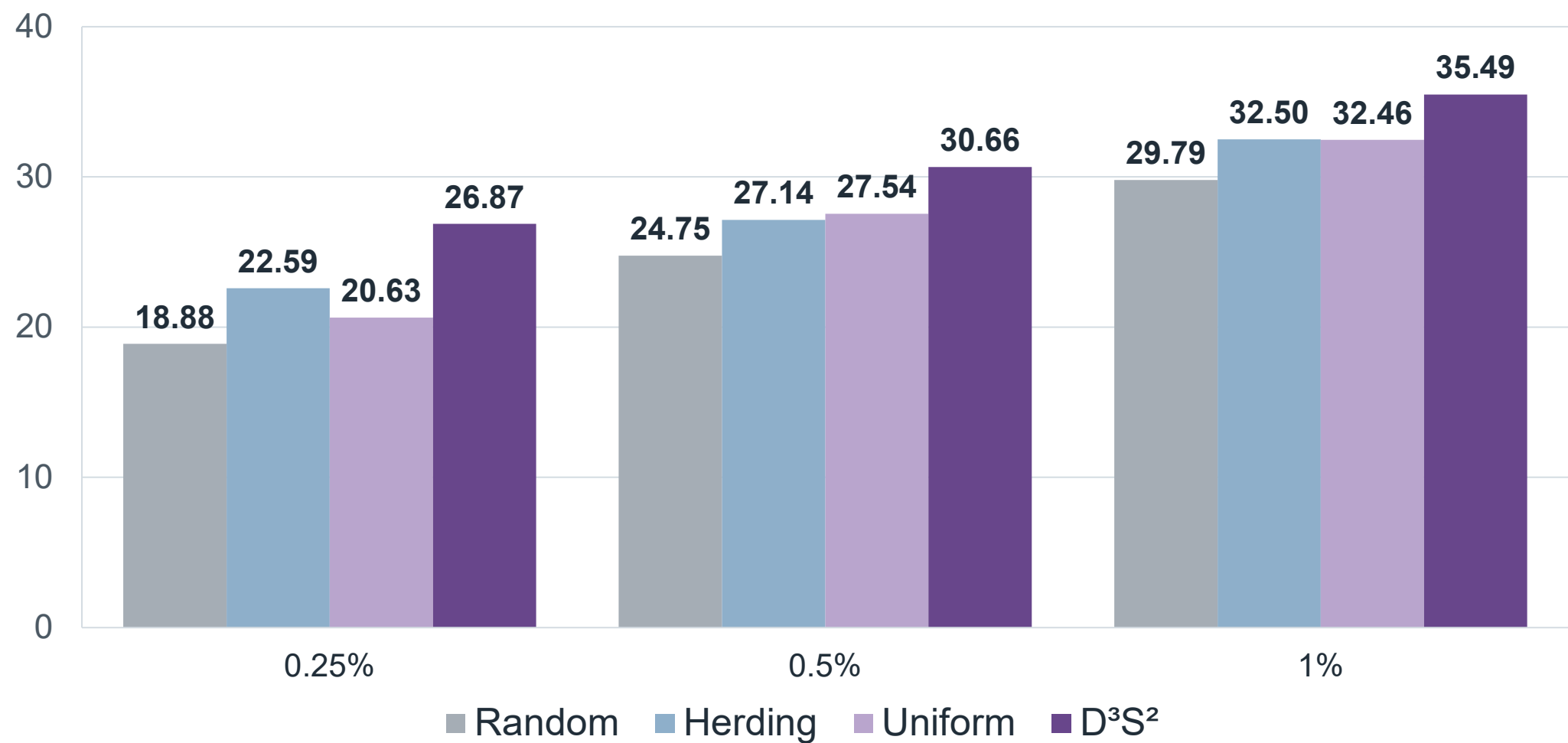
+4.69

points over Uniform
with SegFormer (MiT-B4)

The advantage persists across all three ADE20K budgets (0.5%, 1%, 2%).

COCO-Stuff: gains persist across data budgets

Mask2Former (Swin-S). Mean mIoU (%).



35.49

mIoU at 1% data

+5.70

points over Random

Full-data training reaches 48.19 mIoU. A substantial performance gap remains.

At 0.25%, D³S² reaches **26.87 mIoU** with only **295 distilled images**.

Ablation: inversion and guidance improve utility

ADE20K at 1% data. Mask2Former (Swin-T).

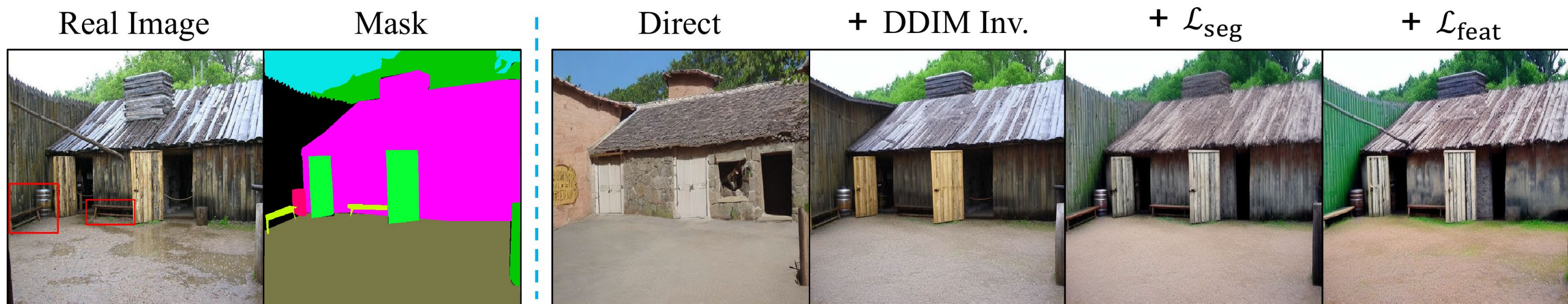
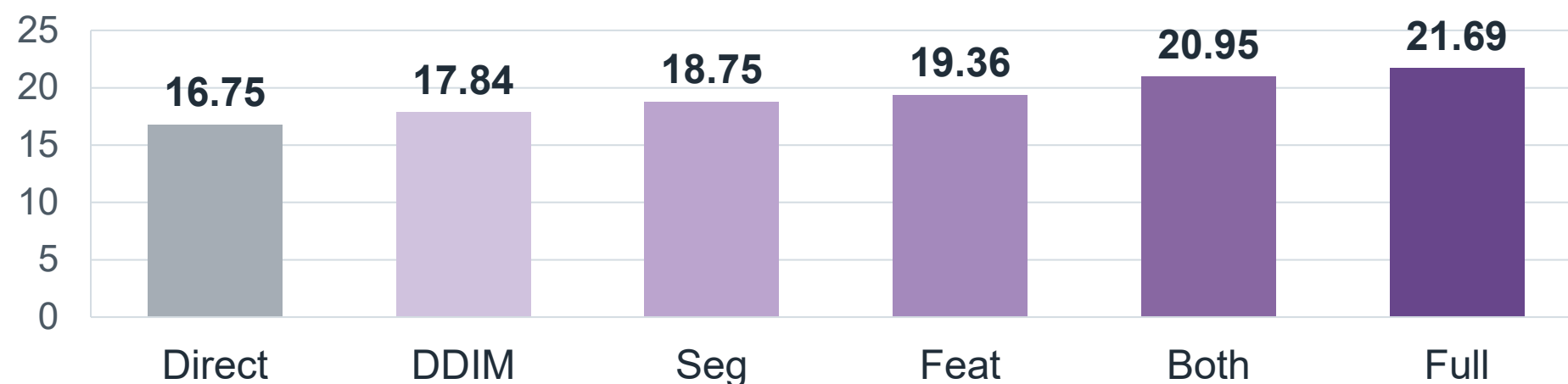


Figure 4. Inversion preserves small objects, while guidance improves geometry and appearance



Complementary guidance

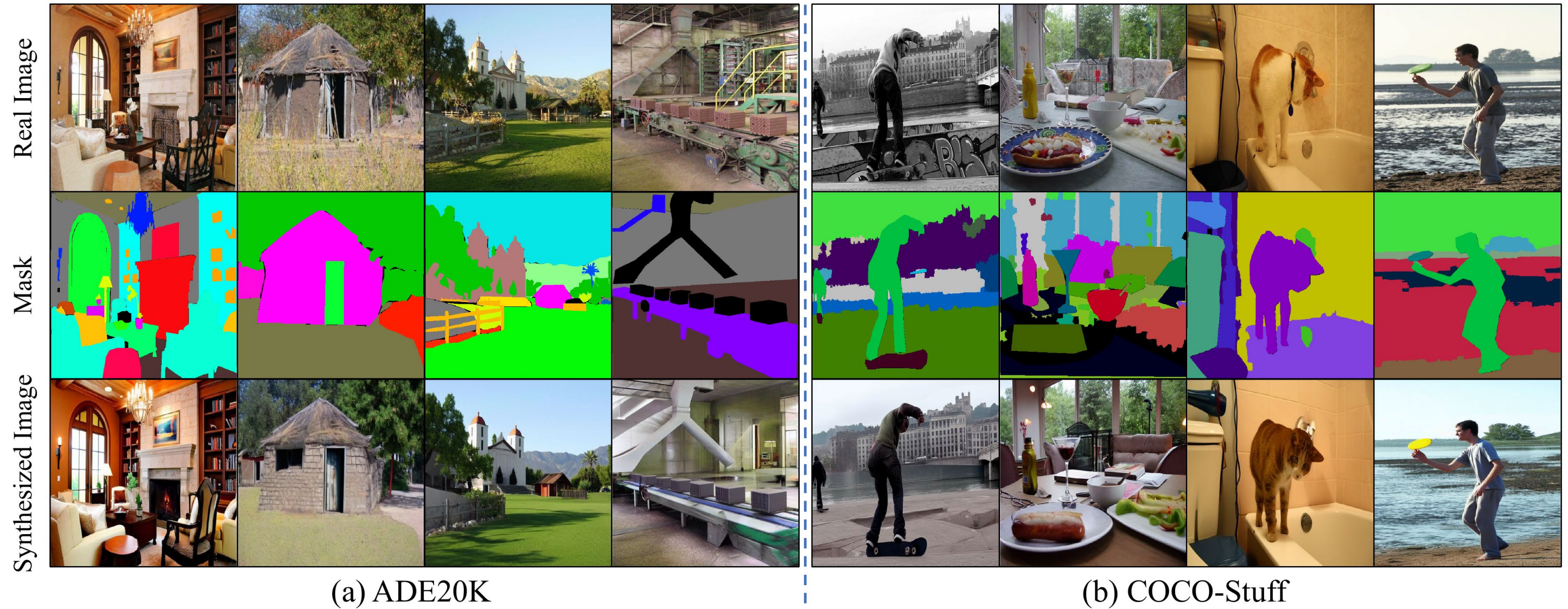
Both losses reach **20.95 mIoU**, compared with 18.75 for Seg and 19.36 for Feat.

Relabeling adds **0.74 points**.

Seg and Feat each add one loss to DDIM. Full adds relabeling to Both. Bars show mean mIoU (%).

Layout fidelity and class-specific appearance

Synthesized scenes retain object boundaries and clarify semantic regions in both datasets.



Rare-class utility: conveyor-belt IoU improves from **11.33** (Uniform) to **30.13** at 2% ADE20K (Swin-T).

Remaining limitation: fine details such as human hands may contain visible artifacts.

Conclusions and resources

Class coverage

Greedy selection retains rare classes within a small mask budget.

Training utility

Pixel-level consistency and class-wise means guide diffusion toward useful segmentation samples.

Practical synthesis

50 denoising steps synthesize each image in about 50 s. Pretrained-model training is an additional cost.



Paper



Code

arxiv.org/abs/2605.25022

github.com/zwj084/D3S2

Contact

Wenjie Zheng (郑文捷)

wenjie_zheng@zju.edu.cn

Haoji Hu (胡浩基)

haoji_hu@zju.edu.cn