Structure-Aware Representation Distillation for Tiny-Dense Object Segmentation

Supplementary Material

6. Evaluation on Dense Segmentation Datasets

To demonstrate the generalizability of SARD beyond the RockFrag, Cityscapes [1], and ADE20K [6] datasets evaluated in the main paper, we conduct experiments on four additional dense-object datasets: MoNuSeg [2], iSAID [5], CellBin [3], and LoveDA [4]. These datasets span medical imaging, aerial imagery, microscopy, and remote sensing, and exhibit extreme object density, large scale variation, and complex boundaries, which are precisely the scenarios where structure-aware distillation is expected to be particularly effective.

6.1. Dataset Descriptions

We evaluate SARD on four additional dense-object datasets spanning different domains:

- MoNuSeg: 30 training and 14 test H&E histopathology images with ∼22K annotated nuclei for nuclear instance segmentation.
- iSAID: 2,806 high-resolution aerial images with 655,451 object instances across 15 categories for instance segmentation.
- CellBin: 1,044 multimodal microscopy images (four staining modalities) with over 109K annotated cells.
- LoveDA: 5,987 0.3 m-resolution remote-sensing images from urban and rural regions, annotated with seven land-cover classes.

6.2. Experimental Setup

All experiments on the four dense datasets reuse the same architecture, teacher–student pairs, input resolution, and hyperparameters as in Sec. 3 (RockFrag, Cityscapes, ADE20K); no additional tuning is performed.

6.3. Results and Analysis

Tables 6 and 7 report results on the four dense-object datasets for both teacher–student pairs.

Consistent improvements across datasets. Across all four datasets and both teacher–student pairs, SARD achieves consistent gains over the baselines. Relative to the strongest baseline (typically CWD), SARD improves both mIoU and bIoU, with the largest gains on the most densely populated datasets (CellBin and MoNuSeg), where fine boundaries and crowded regions are most prevalent.

Enhanced boundary quality. The gains in boundary IoU are particularly notable: SARD tends to yield larger improvements in bIoU than in mIoU, indicating that the student benefits especially in boundary regions. For example, on MoNuSeg with Swin-L → Swin-T, SARD improves mIoU from 79.8 to 82.7 and bIoU from 74.1 to 76.8 compared to CWD (Table 6), suggesting more accurate boundary representations in dense nuclear regions.

Domain-specific observations. On iSAID (aerial imagery), the improvements are more moderate than on microscopy datasets, likely because aerial scenes contain a mixture of large structures (buildings, harbors) and small dense objects (vehicles). In such heterogeneous settings, structure-aware weighting appears to emphasize boundary-rich regions without harming performance on large, relatively homogeneous areas. On LoveDA, SARD improves both urban and rural scenes, indicating that the same structural cues can be applied across domains with different object scales and layout patterns.

7. Structure-Aware Representation Weighting

This section expands the formulation in Sec. 3 by detailing how we compute the structure score $S(i)$ and how the resulting weights $W(i)$ reshape the representation loss in Eq. (4).

7.1. Structure Tensor and Geometric Measures

Given teacher features $F_T \in \mathbb{R}^{C\times H\times W}$ from Eq. (1), we view *structure* as a spatially varying geometric descriptor that captures local organization in the feature field.

For each channel c and location $i \in \Omega$, we compute spatial gradients

$$\nabla F_T^{(c)}(i) = \begin{bmatrix} \frac{\partial F_T^{(c)}}{\partial x}(i) \\ \frac{\partial F_T^{(c)}}{\partial y}(i) \end{bmatrix} \in \mathbb{R}^2.$$

To aggregate directional information across channels, we form the (feature-space) structure tensor

$$\mathbf{J}(i) = \sum_{c=1}^{C} \nabla F_T^{(c)}(i)\, \nabla F_T^{(c)}(i)^\top \in \mathbb{R}^{2\times 2}, \qquad (16)$$

a 2×2 symmetric matrix encoding gradient energy and orientation.

Table 6. Comparison of knowledge distillation methods with Swin-L → Swin-T on dense segmentation datasets.

	MoNuSeg [2]		iSAID [5]		CellBin [3]		LoveDA [4]	
Method	mIoU↑	bIoU↑	mIoU↑	bIoU↑	mIoU↑	bIoU↑	mIoU↑	bIoU↑
Teacher	85.2	78.9	68.4	64.1	82.7	77.3	76.8	71.5
Student (scratch)	76.3	70.1	59.8	55.7	74.1	68.8	68.9	63.2
Vanilla KD	78.5	72.3	62.2	58.1	76.6	71.1	71.3	65.8
CWD	79.8	74.1	63.9	59.8	78.2	72.7	73.1	67.4
DKD	80.1	74.6	64.3	60.2	78.8	73.2	73.6	68.1
MGD	81.3	75.2	65.1	61.1	79.4	74.1	74.2	68.9
SARD (ours)	82.7	76.8	66.5	62.7	80.9	75.6	75.8	70.3
Δ vs CWD	+2.9	+2.7	+2.6	+2.9	+2.7	+2.9	+2.7	+2.9

Table 7. Comparison of knowledge distillation methods with Swin-B → Swin-S on dense segmentation datasets.

	MoNuSeg [2]		iSAID [5]		CellBin [3]		LoveDA [4]	
Method	mIoU↑	bIoU↑	mIoU↑	bIoU↑	mIoU↑	bIoU↑	mIoU↑	bIoU↑
Teacher	81.2	74.9	64.4	60.1	78.7	73.3	72.8	67.5
Student (scratch)	78.8	72.6	62.3	58.2	76.6	71.3	71.4	65.7
Vanilla KD	79.3	73.1	62.6	58.5	76.9	71.6	71.7	66.0
CWD	79.8	73.6	62.9	58.8	77.2	71.9	72.1	66.4
DKD	80.0	73.8	63.1	59.0	77.4	72.1	72.3	66.6
MGD	80.3	74.1	63.3	59.3	77.6	72.4	72.5	66.9
SARD (ours)	80.9	74.8	63.8	59.9	78.1	73.0	73.0	67.5
Δ vs CWD	+1.1	+1.2	+0.9	+1.1	+0.9	+1.1	+0.9	+1.1

Since $\mathbf{J}(i)$ is symmetric, it admits an eigendecomposition

$$\mathbf{J}(i) = \lambda_1(i)\,\mathbf{v}_1(i)\mathbf{v}_1(i)^\top + \lambda_2(i)\,\mathbf{v}_2(i)\mathbf{v}_2(i)^\top, \quad (17)$$

where $\lambda_1(i) \geq \lambda_2(i) \geq 0$ are the principal gradient magnitudes and $\mathbf{v}_1(i), \mathbf{v}_2(i)$ the corresponding directions. Their relative magnitudes characterize standard structure types:

$$\begin{aligned} \text{edge / boundary:} &\quad \lambda_1(i) \gg \lambda_2(i) \approx 0, \\ \text{corner / junction:} &\quad \lambda_1(i) \approx \lambda_2(i) \gg 0, \\ \text{homogeneous:} &\quad \lambda_1(i) \approx \lambda_2(i) \approx 0. \end{aligned} \quad (18)$$

Following the main text, we summarize this behavior with two scalar geometric measures:

$$E(i) = \lambda_1(i) - \lambda_2(i), \qquad C(i) = \sqrt{\lambda_1(i)\lambda_2(i)}, \quad (19)$$

where $E(i)$ captures anisotropy (strong oriented transitions such as edges and contact boundaries) and $C(i)$ captures multi-directional structure (corners, junctions, and complex surface variations). These quantities directly instantiate the geometric terms in Eq. (9).

In practice, spatial gradients $\nabla F_T^{(c)}$ are implemented with standard Sobel filters, and λ_1, λ_2 are computed in closed form for 2×2 matrices, introducing negligible overhead relative to the backbone forward pass.

7.2. Multi-Scale Structure and Density Cues

Multi-scale structure. Tiny-dense objects often contain both fine edges (e.g., fragment chips, thin membranes) and coarser boundaries (e.g., large fragment outlines, major tissue interfaces). To capture this spectrum, we compute structure at multiple spatial scales. Let G_σ denote a Gaussian kernel with standard deviation σ. We define a smoothed feature map $F_{T,\sigma}^{(c)} = G_\sigma * F_T^{(c)}$ and corresponding structure tensor

$$\mathbf{J}_\sigma(i) = \sum_{c=1}^{C} \nabla F_{T,\sigma}^{(c)}(i)\, \nabla F_{T,\sigma}^{(c)}(i)^\top. \quad (20)$$

From $\mathbf{J}_\sigma(i)$ we obtain eigenvalues $\lambda_{1,\sigma}(i), \lambda_{2,\sigma}(i)$ and an anisotropy score

$$E_\sigma(i) = \lambda_{1,\sigma}(i) - \lambda_{2,\sigma}(i).$$

We then aggregate edge evidence across a small pyramid of scales, e.g. $\sigma \in \{1, 2, 4, 8\}$ pixels:

$$E(i) = \sum_{\sigma} w_\sigma E_\sigma(i), \qquad w_\sigma = \exp(-\sigma/\sigma_0). \tag{21}$$

For notational simplicity we reuse $E(i)$ to denote this multi-scale edge strength in Eq. (9).

Density via feature dispersion. Geometric cues alone do not distinguish isolated edges from boundaries inside crowded regions. To emphasize where many objects cluster (e.g., gravel piles or dense cell fields), we introduce a density term $D(i)$ based on local feature dispersion.

For a square window $\mathcal{W}_r(i)$ centered at i, we define the local teacher-feature mean

$$\mu(i) = \frac{1}{|\mathcal{W}_r(i)|} \sum_{j \in \mathcal{W}_r(i)} F_T(j)$$

and measure dispersion as

$$D(i) = \frac{1}{|\mathcal{W}_r(i)|} \sum_{j \in \mathcal{W}_r(i)} \big\| F_T(j) - \mu(i) \big\|_2. \tag{22}$$

High $D(i)$ indicates many distinct feature patterns co-occurring in a small neighborhood, a hallmark of tiny-dense regions where precise discrimination between nearby instances is required. As stated in Sec. 3.2, we use this feature-dispersion definition of $D(i)$ for both labeled and unlabeled images to keep SARD compatible with fully and semi-supervised settings.

7.3. From Structure Score $S(i)$ to Weights $W(i)$

The structure score $S(i)$ in Eq. (9) combines geometric complexity and density:

$$S(i) = \beta_e E(i) + \beta_c C(i) + \beta_d D(i), \tag{23}$$

where $\beta_e, \beta_c, \beta_d \geq 0$ control the relative contributions of edge strength, corner/junction complexity, and spatial crowding. Intuitively:

- $E(i)$ highlights oriented discontinuities such as boundaries between fragments or tissue interfaces.
- $C(i)$ emphasizes multi-directional structures such as T-junctions, corners, and complex facets.
- $D(i)$ up-weights regions where many heterogeneous instances or textures overlap.

To convert $S(i)$ into spatial weights, we apply the normalization in Eq. (5):

$$W(i) = \frac{S(i)}{\sum_{j \in \Omega} S(j)}, \qquad \sum_{i \in \Omega} W(i) = 1. \tag{24}$$

These weights $W(i)$ directly modulate the representation loss $\mathcal{L}_{\text{repr}}$ in Eq. (4) and the feature consistency and distribution-alignment components in Eq. (10), concentrating learning signals on structurally important locations.

7.4. Effect on Representation Loss

We now clarify how $W(i)$ reshapes gradients compared to uniform feature matching.

Uniform feature consistency. A standard (non-structure-aware) feature consistency objective aligns projected features uniformly:

$$\mathcal{L}_{\text{FC}}^{\text{uni}} = \sum_{i \in \Omega} \big\| \hat{F}_S(i) - \hat{F}_T(i) \big\|_2^2, \tag{25}$$

where $\hat{F}_T = P_T(F_T)$ and $\hat{F}_S = P_S(F_S)$ are the projected teacher and student features. The gradient with respect to $\hat{F}_S(i)$ is

$$\frac{\partial \mathcal{L}_{\text{FC}}^{\text{uni}}}{\partial \hat{F}_S(i)} = 2\big(\hat{F}_S(i) - \hat{F}_T(i)\big),$$

so all locations in Ω are treated equally, regardless of whether they lie in flat background or complex boundaries. In tiny-dense datasets, this dilutes the influence of critical boundaries and crowded regions.

Structure-weighted feature consistency. SARD replaces Eq. (25) with the structure-weighted objective in Eq. (11):

$$\mathcal{L}_{\text{FC}} = \sum_{i \in \Omega} W(i) \big\| \hat{F}_S(i) - \hat{F}_T(i) \big\|_2^2, \tag{26}$$

whose gradient becomes

$$\frac{\partial \mathcal{L}_{\text{FC}}}{\partial \hat{F}_S(i)} = 2\, W(i)\big(\hat{F}_S(i) - \hat{F}_T(i)\big).$$

Because $W(i)$ is large where $E(i)$, $C(i)$, or $D(i)$ are high, mismatches at boundaries, junctions, and dense clusters receive much stronger gradients than mismatches in homogeneous regions. In effect, SARD *redistributes* the distillation signal from "easy" background pixels to structurally informative locations, shifting the learned student feature space toward tiny-dense object characteristics without changing the architecture or inference-time cost.

This same weighting applies to the distribution-alignment loss $\mathcal{L}_{\text{DA}}$ in Eq. (13), so both pointwise feature alignment and local distribution preservation are driven most strongly by structurally important regions.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 2, 5, 1

[2] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE TMI*, 36(7):1550–1560, 2017. 1, 2

[3] Can Shi, Jinghong Fan, Zhonghan Deng, Huanlin Liu, Qiang Kang, Yumei Li, Jing Guo, Jingwen Wang, Jinjiang Gong, Sha Liao, Ao Chen, Ying Zhang, and Mei Li. Cellbindb: A large-scale multimodal annotated dataset for cell segmentation with benchmarking of universal models. *GigaScience*, 14:giaf069, 2025. 1, 2

[4] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *NeurIPS Track on Datasets and Benchmarks*, 2021. 1, 2

[5] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *CVPRW*, pages 28–37, 2019. 1, 2

[6] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 2, 5, 1