TYPE-AWARE REPRESENTATION LEARNING WITH DUAL-TEACHER CONSISTENCY FOR LABEL-EFFICIENT NUCLEI SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Nuclei instance segmentation in histopathology requires precise separation of touching nuclei and accurate cell-type classification. However, dense annotations are costly, and compact networks must preserve fine structural details while learning discriminative semantic representations. To address these challenges, we propose SPiN, a compact multitask architecture, together with dual-teacher Mean Teacher training (DTMT) for image-level semi-supervised nuclei instance segmentation. SPiN integrates a FastViT encoder, a parallel local enhancement pathway, and a shared decoder, with auxiliary type supervision to strengthen semantic representations. DTMT employs two exponential moving average teachers with distinct update rates to capture complementary temporal information. Their type predictions are combined through confidence-weighted fusion and filtered to provide soft supervision for unlabeled images, while geometric displacement regression remains supervised by labeled instances. Only the student network is required at inference. On PanNuke, the framework uses 12.50 million parameters and achieves bPQ/mPQ scores of 0.558/0.396 with 10% labeled training images, compared with 0.488/0.304 for supervised-only training and 0.553/0.390 for an adapted UniMatch baseline. These results demonstrate that integrating compact multitask modeling with type-focused temporal consistency improves annotation efficiency while maintaining a compact inference model, providing an effective approach to nuclei instance segmentation under limited supervision.

1 INTRODUCTION

Nuclei instance segmentation connects tissue appearance to quantitative descriptions of cellular organization. A segmentation model must identify each nucleus, separate touching instances, and assign a biologically meaningful type. These tasks depend on both fine spatial detail and discriminative semantic representations. Public benchmarks such as PanNuke and Lizard provide instance masks and cell-type annotations across diverse tissues, enabling systematic evaluation of joint segmentation and classification (Gamper et al., 2019; Graham et al., 2021). Additional training images require costly instance delineation and type assignment.

Learning from fewer annotated images therefore requires effective use of the information available in both labeled and unlabeled tissue. At the same time, compact networks must allocate their representational capacity between local separation cues and category information. HoVer-style models express this division through nuclear foreground, horizontal–vertical (HV) displacement, and type prediction (Graham et al., 2019). HV fields describe instance-relative geometry, whereas type probabilities describe pixel-level semantic categories.

We address this setting with SPiN and dual-teacher Mean Teacher training (DTMT). SPiN combines a compact shared decoder with local feature enhancement and auxiliary type supervision. DTMT then transfers temporally averaged type predictions from two teachers to the student. Type consistency extends semantic learning to unlabeled images while geometric supervision uses annotated instances. Only the student is required at inference.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

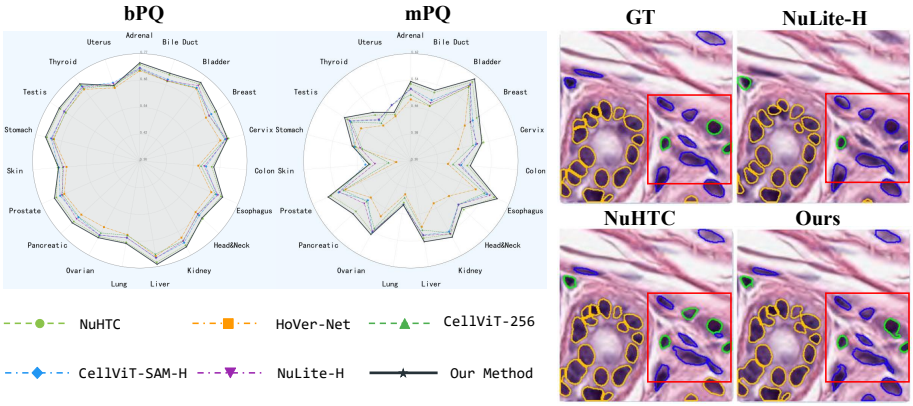

Figure 1: PanNuke comparison across tissues and representative nuclei. Left: bPQ and mPQ radar plots for six methods over 19 tissues. Right: GT and predictions from NuLite-H, NuHTC and Ours (SPiN); red boxes mark the same representative area.

Figure 1 illustrates the resulting segmentation performance. Building on the FastViT-based NuLite architecture (Tommasino et al., 2026), we reinforce spatial detail through a local-aware enhancement unit (LAEU) and category information through a type-aware semantic path (TASP). Two EMA teachers provide complementary temporal targets, which are fused according to confidence. The resulting model achieves 0.558 bPQ and 0.396 mPQ with only 10% labeled training images while retaining a 12.50M-parameter inference model.

Our contributions are threefold: (i) a compact type-aware architecture that combines enhanced local features and auxiliary semantic supervision with shared decoding; (ii) a dual-teacher consistency strategy that integrates two temporal scales into confidence-filtered supervision of the type branch; and (iii) an evaluation of annotation efficiency, component contributions and computational cost, including full label-budget sweeps on PanNuke and Lizard.

2 Related Work

Nuclei instance segmentation. Detection-based, distance-regression, star-convex and microscopy-specific models provide complementary strategies for separating individual nuclei (He et al., 2017; Naylor et al., 2019; Raza et al., 2019; Schmidt et al., 2018; Stringer et al., 2021). HoVer-Net jointly predicts foreground, HV displacements and nuclear types (Graham et al., 2019), while CellViT introduces transformer representations and NuHTC couples detection and segmentation through a hybrid task cascade (Hörst et al., 2024; Li et al., 2025). SPiN instead strengthens the shared representation supporting both instance structure and type prediction.

Compact dense representations. FastViT uses structural reparameterization for efficient visual representations, and MobileViT combines convolutional locality with transformer context (Vasu et al., 2023; Mehta & Rastegari, 2022). NuLite applies FastViT and shared decoding to nuclei segmentation; HoVer-UNet and DualU-Net likewise target efficient instance prediction (Tommasino et al., 2026; 2024; Anglada-Rotger et al., 2025). SPiN preserves this compact shared-decoder design while adding local enhancement and auxiliary type supervision.

Semi-supervised segmentation. Mean Teacher transfers EMA predictions to a student (Tarvainen & Valpola, 2017). Consistency under transformations, features and tasks has been explored in generic and medical segmentation (Bortsova et al., 2019; Ouali et al., 2020; French et al., 2020; Yu et al., 2019; Luo et al., 2021; Wu et al., 2022). FixMatch, CPS, PseudoSeg, ST++, UniMatch and BCP exploit pseudo-labels or strong perturbations (Sohn et al., 2020; Chen et al., 2021; Zou et al., 2021; Yang et al., 2022; 2023; Bai et al., 2023). BASS and UniMatch V2 strengthen boundary-aware or semantic representations (Zhang et al., 2024; Yang et al., 2025). DTMT instead fuses two EMA time scales and targets nuclear type probabilities.

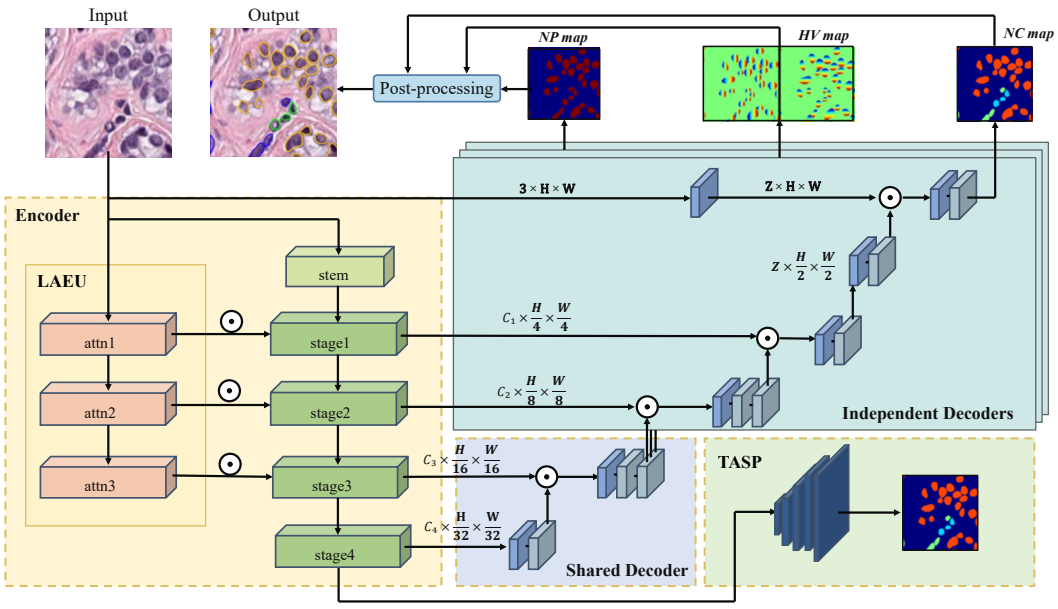

Figure 2: Overall SPiN pipeline. The FastViT encoder extracts multiscale features; the parallel LAEU path supplies local information to the skip connections. A shared decoder supports nuclear foreground (NP), horizontal–vertical displacement (HV), and nuclear category (NC, denoted TP in the text) prediction. TASP supplies auxiliary type supervision. C_1–C_4 denote stage channels. Foreground and HV recover instances; TP assigns their types.

3 METHOD

3.1 PROBLEM SETTING AND FRAMEWORK OVERVIEW

We consider image-level semi-supervised learning within a dataset. The labeled set is $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ and the unlabeled set is $\mathcal{D}_u = \{x_j\}_{j=1}^{N_u}$. The annotation budget is specified by $r = N_u/(N_l + N_u)$, the fraction of training images whose complete annotations are hidden. Thus $r = 0.9$ denotes 10% labeled images. Validation and test sets are held separately from this training budget.

For an image $x \in \mathbb{R}^{H\times W\times 3}$, SPiN predicts foreground $\hat{y}^{\mathrm{np}}$, a two-channel displacement field $\hat{y}^{\mathrm{hv}}$, and type logits $\hat{y}^{\mathrm{tp}}$ over C nuclear types plus background. The auxiliary output $\hat{y}^{\mathrm{aux}}$ provides additional type supervision. Figure 2 presents the student architecture, while Figure 3 describes its teacher-guided training. The framework couples spatial and semantic learning through shared features and uses type consistency to exploit unlabeled images.

3.2 SHARED DECODING AND TYPE-AWARE SUPERVISION

The FastViT encoder produces four feature scales, $(z_1, z_2, z_3, z_4) = E(x)$. SPiN follows the NuLite shared-decoder design (Tommasino et al., 2026): shallow features support spatial recovery and the bottleneck supplies semantic context. The main decoder integrates locally enhanced skips z_1', z_2', z_3' with z_4 to form $f_{\mathrm{dec}} = D(z_1', z_2', z_3', z_4)$. Three lightweight heads then compute the NP, HV and TP outputs from f_{dec}. Sharing the high-resolution recovery path allows the tasks to use a common representation without replicating a complete decoder for each output.

TASP adds an independent auxiliary type path from the bottleneck, $\hat{y}^{\mathrm{aux}} = h_{\mathrm{aux}}(z_4)$. Its supervision directly encourages the encoder to retain category information before shared decoding. TASP influences the main predictions through the encoder parameters; the final instance type is assigned by the main TP head. This arrangement distinguishes semantic supervision during training from the outputs required for instance recovery.

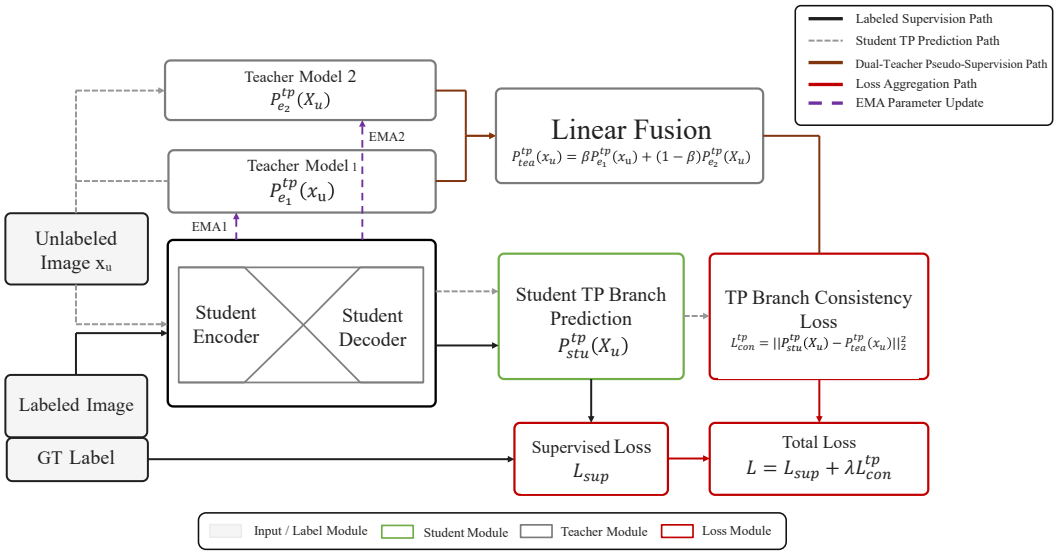

Figure 3: Dual-teacher training with type-aware supervision. Labeled images supervise the student through ground-truth targets; unlabeled images receive pseudo-supervision from two EMA teachers. The teacher type distributions are combined using confidence-weighted fusion and filtered before consistency learning. Only the student receives gradient updates. EMA updates maintain the teachers at distinct temporal scales.

Supervised objective. On labeled images, the foreground loss combines Dice and focal Tversky terms, the HV loss combines mean squared error and spatial-gradient error, the type loss combines cross-entropy, Dice and multiclass focal Tversky terms, and the auxiliary path uses cross-entropy. With branch weights λ_b, the minibatch objective is

$$\mathcal{L}_{\text{sup}}^{\mathcal{B}} = \frac{\sum_{i\in\mathcal{B}} m_i \left(\lambda_{\text{np}}\mathcal{L}_{\text{np},i} + \lambda_{\text{hv}}\mathcal{L}_{\text{hv},i} + \lambda_{\text{tp}}\mathcal{L}_{\text{tp},i} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux},i}\right)}{\max(1, \sum_{i\in\mathcal{B}} m_i)}, \tag{1}$$

where m_i indicates an annotated image. All supervised branches use this mask, including the auxiliary branch. Each loss therefore contributes only when its ground-truth target is available.

3.3 Dual-teacher temporal consistency

DTMT maintains the student parameters θ_s and two EMA teachers with identical architectures. After student update t, teacher k is updated as $\theta_{e,k}^{(t)} = \alpha_k \theta_{e,k}^{(t-1)} + (1-\alpha_k)\theta_s^{(t)}$. We use $\alpha_1 = 0.99$ and $\alpha_2 = 0.995$. The first teacher tracks recent updates more closely, while the second averages over a longer history. Both contribute to the type target in Figure 3.

For pixel v in an unlabeled image, let $p_k(v)$ be teacher k's type distribution and $a_k(v) = \max_c p_{k,c}(v)$ its confidence. The fusion assigns greater weight to the more confident distribution:

$$\beta(v) = \frac{a_1(v)}{a_1(v) + a_2(v) + \epsilon}, \qquad \bar{p}(v) = \beta(v)p_1(v) + (1-\beta(v))p_2(v). \tag{2}$$

We retain pixels with $q(v) = \mathbf{1}[\max_c \bar{p}_c(v) > \tau]$ and set $\tau = 0.7$. The student matches the detached soft target using

$$\mathcal{L}_{\text{con}}(x) = \frac{\sum_v q(v)\|p_s(v) - \text{stopgrad}(\bar{p}(v))\|_2^2}{\sum_v q(v) + \epsilon}. \tag{3}$$

Soft distributions retain relative support for the nuclear categories, while the confidence mask controls which pixels contribute. The consistency gradients update the shared representation through the TP branch. HV regression uses labeled instances because its displacement targets depend on instance geometry, including the relative position of a pixel within its own nucleus.

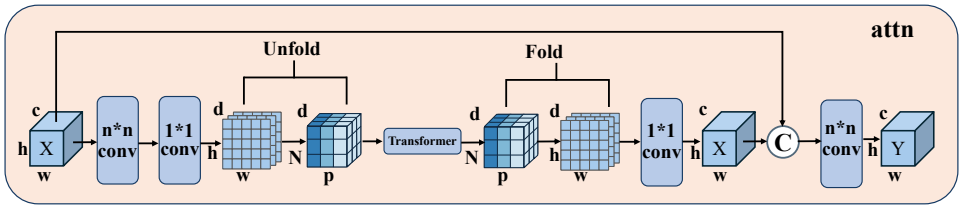

Figure 4: Local-aware enhancement unit. Convolutional processing preserves local spatial information, and the unfold–transformer–fold sequence incorporates contextual information. Channel projection and feature fusion produce the enhanced representation used by the main encoder–decoder pathway.

The total objective combines labeled supervision and type consistency:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}^{\mathcal{B}} + \frac{w(t)}{|\mathcal{B}_u|} \sum_{x \in \mathcal{B}_u} \mathcal{L}_{\text{con}}(x), \quad w(t) = 0.02 \exp\left[-5\left(1 - \min(t/2000, 1)\right)^2\right], \tag{4}$$

An empty unlabeled batch contributes zero consistency loss. The ramp-up gradually introduces teacher supervision.

Confidence weighting and confidence filtering play distinct roles in this objective. The fusion coefficient balances the two teacher distributions at each pixel, whereas the mask determines whether the resulting target contributes to learning. Equal teacher confidence produces approximately equal fusion weights; otherwise, the more confident prediction receives greater weight. Normalization by the number of retained pixels controls the scale of the consistency term across images with different accepted-pixel counts. The soft target carries category information into the TP head and the shared features, while the stop-gradient operation keeps the teachers outside student backpropagation. After the student update, EMA transfers its parameter changes to both teachers at their respective temporal scales. This sequence connects confidence-based target construction to temporal representation learning.

3.4 Local enhancement and instance recovery

The LAEU pathway complements type supervision with spatial detail (Figure 4). It processes the image in parallel with the main encoder using convolutional and MobileViT operations (Mehta & Rastegari, 2022). Let $d_1 = A_1(x)$ and $d_l = A_l(d_{l-1})$. At each of the first three scales, the skip feature becomes $z_l' = \psi_l[z_l \,\|\, \text{Resize}(d_l; z_l)]$, where $\|$ denotes channel concatenation and ψ_l mixes the aligned features. The decoder combines these features through $u_l = \phi_l[\text{Up}(u_{l+1}) \,\|\, z_l']$. This path directly reinforces spatial recovery.

At inference, foreground support and HV changes guide watershed-based instance reconstruction (Graham et al., 2019). Aggregating TP scores within each instance assigns its type. The teachers and auxiliary type output serve training only.

4 Experiments

4.1 Experimental setup

PanNuke is the primary benchmark and covers five nuclear types across 19 tissues (Gamper et al., 2019). Lizard provides an independent label-budget evaluation (Graham et al., 2021), while CoNSeP tests segmentation quality under a distinct colorectal tissue distribution. All low-label experiments are performed within each dataset: annotations are hidden from a fraction of the training images, those images remain available as unlabeled inputs, and validation and test annotations remain complete. PanNuke folds 0/1/2 are used for training, validation and testing.

We train for 130 epochs with batch size 20 using AdamW, a learning rate of 3×10^{-4}, weight decay 10^{-4} and $(\beta_1, \beta_2) = (0.85, 0.95)$. ExponentialLR uses $\gamma = 0.85$. Training runs on an RTX 3090 with automatic mixed precision and spatial and appearance augmentation. DTMT uses EMA coefficients 0.99/0.995, confidence threshold 0.7, consistency weight 0.02 and a 2000-step ramp-up.

Evaluation metrics. We report binary panoptic quality (bPQ), type-aware mPQ, detection precision, recall and F1, together with class-specific F1 where appropriate. For matched instances, $\mathrm{PQ} = \sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)/(|\mathrm{TP}| + \frac{1}{2}|\mathrm{FP}| + \frac{1}{2}|\mathrm{FN}|)$. bPQ measures instance quality without type distinctions, whereas mPQ also requires correct nuclear types. Parameters, FLOPs and latency quantify inference cost under the same evaluation protocol.

4.2 Label-efficient learning

Table 1 establishes the effect of DTMT across four annotation budgets. DTMT ranks first in every reported metric at every displayed budget. With 70% labeled images ($r = 0.3$), it improves Mean Teacher from 0.623/0.471 to 0.628/0.478 bPQ/mPQ. With only 10% labeled images ($r = 0.9$), the margin widens to 0.558/0.396 versus 0.543/0.378. The monotonic widening of the gap as labels are removed demonstrates that the dual temporal targets are most valuable precisely in the regime the method is designed for.

Table 1: Semi-supervised results at four unlabeled-image fractions r. The labeled fraction is $1 - r$.

Unlabeled r	Method	bPQ ↑	mPQ ↑	F1 ↑	Precision ↑	Recall ↑
0.30	EMGS	0.608	0.446	0.814	0.795	0.834
0.30	Mean Teacher	0.623	0.471	0.823	0.809	0.837
0.30	**DTMT (Ours)**	**0.628**	**0.478**	**0.824**	**0.810**	**0.838**
0.50	EMGS	0.592	0.427	0.803	0.788	0.820
0.50	Mean Teacher	0.608	0.451	0.809	0.795	0.824
0.50	**DTMT (Ours)**	**0.613**	**0.459**	**0.811**	**0.797**	**0.826**
0.70	EMGS	0.560	0.389	0.781	0.764	0.799
0.70	Mean Teacher	0.582	0.421	0.789	0.775	0.804
0.70	**DTMT (Ours)**	**0.591**	**0.433**	**0.793**	**0.778**	**0.808**
0.90	EMGS	0.513	0.335	0.746	0.728	0.766
0.90	Mean Teacher	0.543	0.378	0.758	0.743	0.773
0.90	**DTMT (Ours)**	**0.558**	**0.396**	**0.766**	**0.751**	**0.782**

Table 2 compares stronger task-adapted semi-supervised baselines at the hardest 10% label setting. DTMT achieves 0.558 bPQ, 0.396 mPQ and 0.766 detection F1, exceeding UniMatch-adapted by 0.005, 0.006 and 0.002, respectively, while also leading precision and recall. The result is consistent across all five metrics: confidence-weighted dual-teacher targets provide the strongest unlabeled type supervision among the evaluated strategies.

Table 2: Semi-supervised comparison with 10% labeled images on PanNuke.

Method	bPQ ↑	mPQ ↑	F1 ↑	Precision ↑	Recall ↑
EMGS	0.513	0.335	0.746	0.728	0.766
Mean Teacher	0.543	0.378	0.758	0.743	0.773
FixMatch-adapted	0.529	0.356	0.752	0.738	0.770
CPS-adapted	0.546	0.381	0.759	0.745	0.776
UniMatch-adapted	0.553	0.390	0.764	0.748	0.780
DTMT (Ours)	**0.558**	**0.396**	**0.766**	**0.751**	**0.782**

The complete PanNuke and Lizard sweeps show the same pattern over the full range of label budgets. On PanNuke at 10% labels, DTMT improves supervised-only training by 0.070 bPQ, 0.092 mPQ and 0.058 detection F1. It preserves 86.1% of full-label bPQ and 78.7% of full-label mPQ, compared with 75.3% and 60.4% for supervised-only training. On Lizard at the same budget, DTMT improves bPQ/mPQ/F1 by 0.052/0.060/0.062. These gains confirm that the method consistently converts unlabeled images into measurable instance- and type-level accuracy.

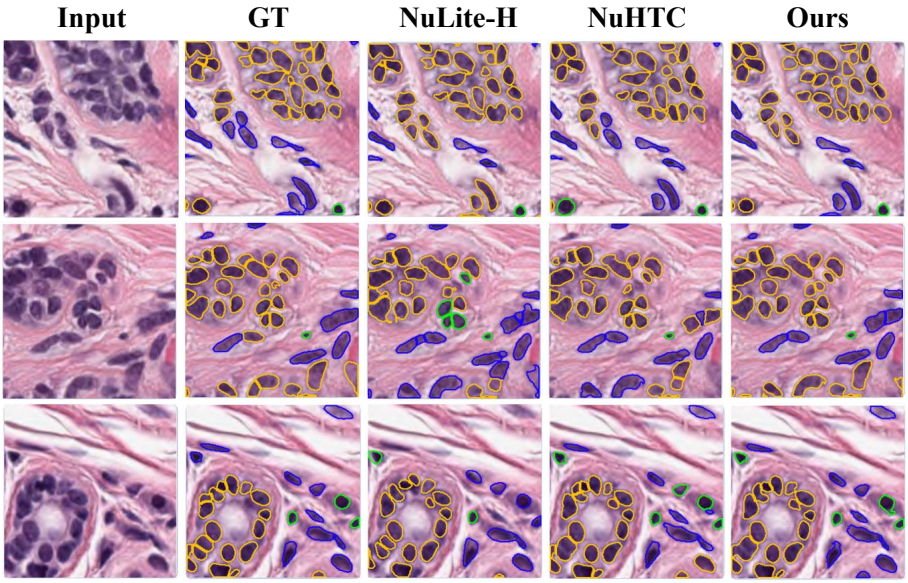

Figure 5: Qualitative nuclei-segmentation comparison: input, ground truth (GT), NuLite-H, NuHTC and Ours (SPiN). Colored contours show instance separation and type assignments in the same regions.

4.3 Full-supervision accuracy and efficiency

Table 3 gives the complete tissue-wise PanNuke comparison. SPiN achieves the strict best mPQ on 14 of 19 tissues and ties the best mPQ on Lung; it also achieves the best bPQ on 15 of 19 tissues. The gains cover diverse organs rather than a narrow subset: for example, SPiN reaches 0.534/0.731 mPQ/bPQ on Adrenal, 0.578/0.697 on Esophagus, 0.556/0.730 on Kidney, and 0.542/0.761 on Liver. The few exceptions are concentrated in Cervix, Head&Neck, Skin and Uterus. This distribution demonstrates broad tissue-level robustness under full supervision.

Table 3: Tissue-wise performance on PanNuke. Each cell reports mPQ ↑ / bPQ ↑.

Tissue	Mask R-CNN	HoVer-Net	CellViT-256	CellViT-SAM-H	NuLite-H	NuHTC	**SPiN (Ours)**
Adrenal	0.347 / 0.555	0.481 / 0.696	0.495 / 0.701	0.513 / 0.709	0.511 / 0.706	0.527 / 0.723	**0.534 / 0.731**
Bile Duct	0.354 / 0.557	0.471 / 0.670	0.472 / 0.671	0.489 / 0.678	0.480 / 0.672	0.513 / 0.700	**0.520 / 0.707**
Bladder	0.507 / 0.605	0.579 / 0.703	0.576 / 0.706	0.584 / 0.707	0.586 / 0.719	0.595 / 0.727	**0.606 / 0.736**
Breast	0.388 / 0.557	0.490 / 0.647	0.509 / 0.664	0.518 / 0.675	0.510 / 0.663	0.532 / 0.693	**0.540 / 0.701**
Cervix	0.340 / 0.548	0.444 / 0.665	0.489 / 0.686	0.498 / 0.687	0.502 / 0.693	**0.521 / 0.702**	0.516 / 0.697
Colon	0.312 / 0.460	0.410 / 0.558	0.425 / 0.570	0.449 / 0.592	0.443 / 0.584	0.467 / 0.617	**0.476 / 0.626**
Esophagus	0.431 / 0.569	0.509 / 0.643	0.537 / 0.662	0.545 / 0.668	0.554 / 0.675	0.568 / 0.689	**0.578 / 0.697**
Head&Neck	0.395 / 0.546	0.453 / 0.633	0.490 / 0.647	0.491 / 0.654	0.491 / 0.645	**0.512 / 0.676**	0.505 / 0.671
Kidney	0.355 / 0.509	0.442 / 0.684	0.541 / 0.699	0.537 / 0.709	0.545 / 0.701	0.546 / 0.722	**0.556 / 0.730**
Liver	0.410 / 0.609	0.497 / 0.725	0.507 / 0.716	0.522 / 0.732	0.523 / 0.733	0.531 / 0.752	**0.542 / 0.761**
Lung	0.318 / 0.513	0.400 / 0.630	0.410 / 0.632	0.431 / 0.643	**0.432** / 0.643	0.425 / 0.663	**0.432 / 0.668**
Ovarian	0.434 / 0.578	0.486 / 0.631	0.526 / 0.660	0.539 / 0.672	0.537 / 0.671	0.539 / 0.677	**0.547 / 0.684**
Pancreatic	0.362 / 0.546	0.460 / 0.649	0.477 / 0.664	0.472 / 0.666	0.486 / 0.677	0.520 / 0.694	**0.531 / 0.703**
Prostate	0.396 / 0.579	0.510 / 0.662	0.516 / 0.670	0.532 / 0.682	0.529 / 0.674	0.557 / 0.700	**0.566 / 0.709**
Skin	0.267 / 0.502	0.343 / 0.623	0.366 / 0.640	**0.434** / 0.657	0.406 / 0.636	0.421 / **0.662**	0.423 / 0.658
Stomach	0.368 / 0.598	0.473 / 0.689	0.448 / 0.692	0.471 / 0.702	0.454 / 0.698	0.467 / 0.721	**0.478 / 0.727**
Testis	0.351 / 0.542	0.475 / 0.689	0.509 / 0.688	0.513 / 0.696	0.517 / 0.697	0.524 / 0.717	**0.533 / 0.723**
Thyroid	0.304 / 0.571	0.432 / 0.698	0.441 / 0.704	0.452 / 0.715	0.454 / 0.710	0.471 / 0.722	**0.481 / 0.729**
Uterus	0.368 / 0.559	0.439 / 0.639	**0.474** / 0.652	**0.474 / 0.663**	0.473 / 0.658	0.444 / 0.644	0.451 / 0.650

SPiN simultaneously attains the highest PanNuke detection F1 (0.84) and CoNSeP Dice (0.847) with only 12.50M parameters, 15.33G FLOPs and 4.923 ms latency (Table 4). Figure 5 further confirms clean instance separation and stable type assignment in challenging regions.

Table 4: Comparison of model efficiency and representative performance on CoNSeP and PanNuke.

	Efficiency			CoNSeP	PanNuke
Model	Params (M) ↓	FLOPs (G) ↓	Latency (ms) ↓	Dice ↑	Det. ↑
CellViT-SAM-H	699.74	942.54	62.799	0.846	0.83
CellViT-256	46.75	277.08	35.710	0.821	0.82
NuLite-H	34.10	23.15	29.990	0.842	0.83
NuHTC	30.75	35.10	40.056	0.818	0.83
SPiN (Ours)	**12.50**	**15.33**	**4.923**	**0.847**	**0.84**

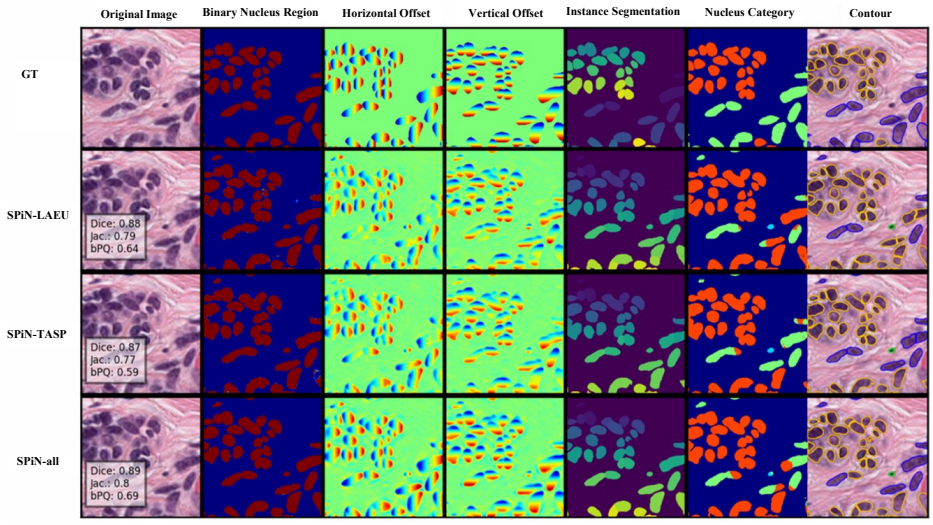

Figure 6: Qualitative architecture ablation on PanNuke. The original image and ground truth are compared with the TASP, LAEU and complete SPiN variants through contours, foreground support, horizontal and vertical offsets, nuclear categories and final instance segmentation.

4.4 Component analysis

Table 5 isolates TASP and LAEU under full supervision. TASP improves F1 over the base model for all five nuclear types, and LAEU produces larger gains by strengthening local structure. Combining both components yields the best F1 in every displayed class. Relative to SPiN-Base, the complete model improves neoplastic, epithelial, inflammatory, connective and dead-cell F1 by 0.020, 0.018, 0.020, 0.019 and 0.028, respectively. The consistent improvement across all categories demonstrates that the two modules are complementary: TASP strengthens category discrimination, while LAEU strengthens boundary and instance structure.

Table 5: Architecture ablation on PanNuke. Base uses shared decoding; TASP adds auxiliary type supervision; LAEU adds local enhancement; All combines both. Each cell lists Precision ↑ / Recall ↑ / F1 ↑.

Class	Base	TASP	LAEU	**All**
Neoplastic	0.669 / 0.707 / 0.688	0.679 / 0.713 / 0.695	0.687 / 0.722 / 0.704	**0.690 / 0.727 / 0.708**
Epithelial	0.690 / 0.717 / 0.703	0.697 / 0.723 / 0.710	0.708 / 0.729 / 0.718	**0.710 / 0.732 / 0.721**
Inflammatory	0.603 / 0.563 / 0.582	0.611 / 0.569 / 0.589	0.619 / 0.579 / 0.598	**0.623 / 0.583 / 0.602**
Connective	0.493 / 0.553 / 0.522	0.502 / 0.559 / 0.528	0.508 / 0.571 / 0.537	**0.511 / 0.575 / 0.541**
Dead	0.356 / 0.412 / 0.382	0.366 / 0.427 / 0.394	0.377 / 0.439 / 0.406	**0.380 / 0.446 / 0.410**

Figure 6 exposes the same complementarity qualitatively: TASP reduces type confusion, LAEU sharpens HV structure and boundaries, and their combination produces the most coherent instance map. Tissue-level ablation is equally consistent: the complete model improves both mPQ and bPQ over SPiN-Base on all 19 tissues, with only isolated single-module wins such as Stomach mPQ.

4.5 COMPLETE ANNOTATION-BUDGET ANALYSIS

Table 6: PanNuke results with DTMT across annotation budgets (r: unlabeled fraction).

Metric / r	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
bPQ	**0.648**	0.640	0.634	0.628	0.621	0.613	0.603	0.591	0.576	0.558
mPQ	**0.503**	0.494	0.486	0.478	0.469	0.459	0.447	0.433	0.416	0.396
Det. F1	**0.840**	0.834	0.829	0.824	0.818	0.811	0.803	0.793	0.781	0.766
PQ_n	**0.548**	0.543	0.538	0.532	0.525	0.517	0.507	0.495	0.480	0.462
PQ_e	**0.537**	0.531	0.526	0.520	0.513	0.505	0.495	0.483	0.468	0.450
PQ_i	**0.472**	0.467	0.462	0.456	0.449	0.441	0.431	0.419	0.404	0.386
PQ_c	**0.438**	0.432	0.426	0.420	0.413	0.405	0.395	0.383	0.368	0.350
PQ_d	**0.032**	0.025	0.020	0.015	0.011	0.008	0.005	0.003	0.002	0.001
$F1_n$	**0.708**	0.702	0.696	0.690	0.683	0.675	0.665	0.653	0.638	0.620
$F1_e$	**0.721**	0.715	0.709	0.703	0.696	0.688	0.678	0.666	0.651	0.633
$F1_i$	**0.602**	0.596	0.590	0.584	0.577	0.569	0.559	0.547	0.532	0.514
$F1_c$	**0.541**	0.534	0.527	0.520	0.512	0.503	0.492	0.479	0.463	0.444
$F1_d$	**0.410**	0.312	0.245	0.184	0.132	0.090	0.041	0.028	0.021	0.016

Table 7: PanNuke results with supervised-only training across annotation budgets.

Metric / r	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
bPQ	**0.648**	0.632	0.619	0.606	0.592	0.576	0.559	0.539	0.516	0.488
mPQ	**0.503**	0.482	0.466	0.449	0.431	0.411	0.389	0.364	0.336	0.304
Det. F1	**0.840**	0.828	0.818	0.807	0.795	0.782	0.767	0.750	0.731	0.708
PQ_n	**0.548**	0.536	0.527	0.517	0.506	0.493	0.479	0.462	0.442	0.418
PQ_e	**0.537**	0.524	0.514	0.503	0.491	0.478	0.463	0.446	0.426	0.402
PQ_i	**0.472**	0.459	0.449	0.438	0.426	0.413	0.399	0.382	0.362	0.338
PQ_c	**0.438**	0.423	0.411	0.398	0.384	0.369	0.353	0.335	0.314	0.289
PQ_d	**0.032**	0.022	0.016	0.011	0.007	0.004	0.002	0.001	0.000	0.000
$F1_n$	**0.708**	0.694	0.683	0.671	0.659	0.645	0.630	0.612	0.592	0.568
$F1_e$	**0.721**	0.707	0.696	0.684	0.671	0.657	0.641	0.623	0.603	0.579
$F1_i$	**0.602**	0.588	0.577	0.565	0.552	0.538	0.523	0.505	0.485	0.461
$F1_c$	**0.541**	0.525	0.512	0.499	0.485	0.469	0.452	0.433	0.411	0.386
$F1_d$	**0.410**	0.287	0.203	0.135	0.072	0.037	0.009	0.000	0.000	0.000

Table 8: Lizard results across unlabeled fractions r. S: supervised-only; D: DTMT.

Method	Metric / r	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
S	bPQ	0.444	0.433	0.423	0.412	0.400	0.386	0.369	0.349	0.325	0.296
S	mPQ	0.390	0.377	0.365	0.352	0.338	0.322	0.302	0.279	0.252	0.221
S	F1	0.714	0.701	0.689	0.676	0.662	0.646	0.627	0.605	0.578	0.545
D	bPQ	**0.446**	**0.440**	**0.434**	**0.428**	**0.421**	**0.413**	**0.401**	**0.387**	**0.370**	**0.348**
D	mPQ	**0.392**	**0.385**	**0.378**	**0.371**	**0.363**	**0.354**	**0.340**	**0.324**	**0.305**	**0.281**
D	F1	**0.715**	**0.709**	**0.703**	**0.696**	**0.688**	**0.679**	**0.666**	**0.650**	**0.631**	**0.607**

With 10% labels, DTMT gains 0.070 bPQ, 0.092 mPQ and 0.058 detection F1 over supervised-only training on PanNuke; the corresponding Lizard gains are 0.052, 0.060 and 0.062. PanNuke class-F1 gains are 0.052/0.054/0.053/0.058/0.016. The dead class remains limiting (0.016 versus 0.410 full-label F1), identifying rare-class annotation coverage as the remaining bottleneck.

5 CONCLUSION

SPiN and DTMT provide a compact solution for label-efficient nuclei instance segmentation: SPiN reaches 0.84 detection F1 with 12.50M parameters, while DTMT reaches 0.558 bPQ/0.396 mPQ with only 10% labeled PanNuke images. Consistent gains on PanNuke and Lizard establish the effectiveness of the proposed architecture and training strategy.

AI USE STATEMENT

OpenAI Codex was used solely to assist with manuscript writing, including language polishing, manuscript organization, and LaTeX formatting. All scientific ideas, methodological designs, experiments, analyses, results, and conclusions were developed and verified by the authors. No AI tools were used for literature discovery, research ideation or execution, synthetic data generation, or mathematical proof.

ETHICS STATEMENT

This study uses public histopathology benchmarks for methodological research in nuclei instance segmentation. The evaluation concerns benchmark segmentation and classification performance. Clinical deployment requires validation for the intended tissue, acquisition conditions and diagnostic workflow, including the effects of instance-separation and rare-type errors on downstream cell statistics.

REPRODUCIBILITY STATEMENT

Section 3 specifies the architecture and learning objectives, and Section 4 summarizes the experimental protocol, label-budget comparisons, efficiency results and component analysis. The editable package contains all LaTeX tables and figure sources needed to reproduce the reported manuscript layout.

REFERENCES

David Anglada-Rotger, Berta Jansat, Ferran Marques, and Montse Pardàs. Two heads are enough: DualU-Net, a fast and efficient architecture for nuclei instance segmentation. In *Medical Imaging with Deep Learning*, 2025. URL https://mlanthology.org/midl/2025/angladarotger2025midl-two/.

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11514–11524, 2023. doi: 10.1109/CVPR52729.2023.01108.

Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, pp. 810–818, 2019. doi: 10.1007/978-3-030-32226-7_90.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021.

Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference*, 2020. doi: 10.5244/C.34.154.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. PanNuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, pp. 11–19, 2019.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. doi: 10.1016/j.media.2019.101563.

Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, Noorul Wahab, Fayyaz Minhas,

Shan E. Ahmed Raza, Hesham El Daly, Kishore Gopalakrishnan, David Snead, and Nasir M. Rajpoot. Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 684–693, 2021.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek. CellViT: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. doi: 10.1016/j.media.2024.103143.

Bao Li, Zhenyu Liu, Song Zhang, Xiangyu Liu, Caixia Sun, Jiangang Liu, Bensheng Qiu, and Jie Tian. NuHTC: A hybrid task cascade for nuclei instance segmentation and classification. *Medical Image Analysis*, 103:103595, 2025. doi: 10.1016/j.media.2025.103595.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8801–8809, 2021. doi: 10.1609/aaai.v35i10.17066.

Sachin Mehta and Mohammad Rastegari. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vh-0sUt8HlG.

Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 38 (2):448–459, 2019. doi: 10.1109/TMI.2018.2865709.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12674–12684, 2020.

Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M. Rajpoot. Micro-Net: A unified model for segmentation of various objects in microscopy images. *Medical Image Analysis*, 52:160–173, 2019. doi: 10.1016/j.media.2018.12.003.

Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, pp. 265–273, 2018. doi: 10.1007/978-3-030-00934-2_30.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods*, 18:100–106, 2021. doi: 10.1038/s41592-020-01018-x.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Cristian Tommasino, Cristiano Russo, Antonio Maria Rinaldi, and Francesco Ciompi. HoVer-UNet: Accelerating HoVerNet with U-Net-based multi-class nuclei segmentation via knowledge distillation. In *IEEE International Symposium on Biomedical Imaging*, pp. 1–4, 2024. doi: 10.1109/ISBI56570.2024.10635755.

Cristian Tommasino, Cristiano Russo, and Antonio M. Rinaldi. NuLite: Lightweight and fast model for nuclei instance segmentation and classification. *Biomedical Signal Processing and Control*, 114:109333, 2026. doi: 10.1016/j.bspc.2025.109333. URL https://github.com/CosmoIknosLab/NuLite.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. FastViT: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5762–5772, 2023.

Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022. doi: 10.1016/j.media.2022.102530.

Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. ST++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4268–4277, 2022.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7236–7246, 2023.

Lihe Yang, Zhen Zhao, and Hengshuang Zhao. UniMatch V2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (4):3031–3048, 2025. doi: 10.1109/TPAMI.2025.3528453. URL https://github.com/LiheYoung/UniMatch-V2.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613, 2019. doi: 10.1007/978-3-030-32245-8_67.

Ye Zhang, Ziyue Wang, Yifeng Wang, Hao Bian, Linghan Cai, Hengrui Li, Lingbo Zhang, and Yongbing Zhang. Boundary-aware contrastive learning for semi-supervised nuclei instance segmentation. In *Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 1851–1861, 2024. URL https://proceedings.mlr.press/v250/zhang24a.html.

Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. PseudoSeg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ubSMQiXzcmI.