This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

CIPS-Net: Text-Instructed Conditional Nucleus Instance Segmentation Network in Histopathology

Nikhileswara Rao Sulake[1,†], Sai Manikanta Eswar Machara[1,†], Sivaji Retta[1], Iyyakutti Iyappan Ganapathi[2], Muhammad Owais[2], and Irfan Hussain[2⋆]

[1] Rajiv Gandhi University of Knowledge Technologies, Nuzvid, Andhra Pradesh, India
[2] Khalifa University Center for Autonomous Robotic Systems (KU-CARS), Khalifa University, UAE
Webpage: https://nikhil-rao20.github.io/cips-net/

Abstract. Nucleus instance segmentation is essential for cancer diagnosis, yet existing methods primarily follow a segment-all-then-classify paradigm lacking mechanisms for selective analysis. We present Conditional Instruction-guided Pathology Segmentation Network (CIPS-Net), a text-conditional framework allowing pathologists to segment specific nucleus types via natural language instructions. Our architecture combines a DINOv2 ViT-B/14 vision encoder with Bio_ClinicalBERT text encoding, fusing features through cross-attention and language-modulated skip connections. A four-head design jointly predicts nucleus presence, distance transforms, instance embeddings, and type classification. To encourage conditional behaviour without additional annotations, we introduce permutation-based training that enumerates all $2^N - 1$ non-empty class subsets per image. On PanNuke, CIPS-Net achieves an mPQ of 0.4846 and a bPQ of 0.6217 (3-fold cross-validation), remaining competitive with unconditional SOTA methods while supporting text-conditional selective segmentation at higher inference speeds (32.3 FPS) and lower computational cost (52.6 G MACs) than the top-performing unconditional baseline. Notably, CIPS-Net yields competitive Dead-class PQ (0.211) and Inflammatory PQ (0.471) among evaluated methods, suggesting that text conditioning may benefit the analysis of rare, diagnostically critical cell types.

Keywords: Nucleus Instance Segmentation · Text-Conditioned Segmentation · Histopathology · Permutation-Based Training · DINOv2.

1 Introduction

Histopathological examination is the gold standard for cancer diagnosis. Within this workflow, multi-class nucleus instance segmentation is critical, as nuclear

⋆ Corresponding author: irfan.hussain@ku.ac.ae
† Equal contribution

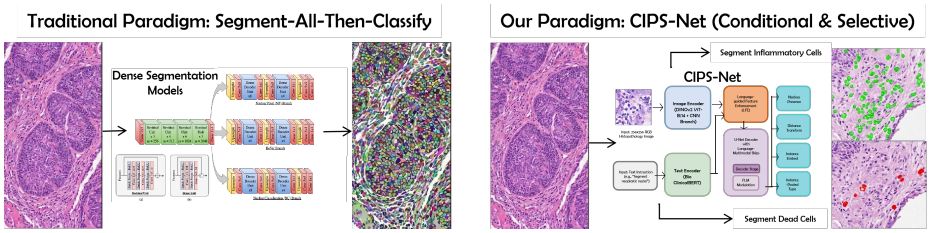

Fig. 1. Traditional "segment-all" approaches vs CIPS-Net's language-conditioned paradigm.

morphology and spatial distribution directly inform tumour grading [7]. However, automated analysis of histopathology images faces considerable challenges such as high density of nuclei, overlapping boundaries, and large variability in nuclear appearance across tissue types and staining conditions [5, 6]. Manual Whole Slide Imaging (WSI) annotation is labor-intensive and also prone to observer variability [8], emphasizing the need for robust automated nucleus instance segmentation. Bottom-up regression approaches dominate this task [7, 18, 4, 14, 10, 20]. These methods achieve strong accuracy but share a fundamental limitation: they follow a *segment-all-then-classify* design that presents all segmented nuclei, forcing pathologists to manually filter irrelevant populations and increasing cognitive load. Pathologists often focus on specific cell populations, yet few methods enable selective, on-demand analysis [24, 21, 27] (see Fig. 1).

While text-conditioned segmentation exists in natural images [25, 23] and tissue-level tasks [21, 1], text-instructed conditional nucleus instance segmentation has not been extensively explored. Bridging this gap enables workflows where pathologists specify targets via natural language, reducing cognitive load and tying outputs directly to human-understandable instructions. Additionally, self-supervised models like DINOv2 [15] offer powerful visual features that could benefit this task.

We propose the Conditional Instruction-guided Pathology Segmentation (CIPS-Net) Network, a text-instructed framework for conditional nucleus instance segmentation. Given an instruction (e.g., "*segment inflammatory cells*"), CIPS-Net integrates textual guidance to segment exclusively the specified types, offering an alternative to post-hoc filtering of "segment-all" outputs. Our key contributions are:

1. We introduce text-instructed conditional segmentation, aiming to support a human-centric workflow that reduces cognitive load by focusing on clinically relevant nuclei.
2. We propose a permutation-based training strategy enumerating all $2^N - 1$ non-empty class subsets, encouraging the model to avoid unrequested segmentations without requiring additional annotations.

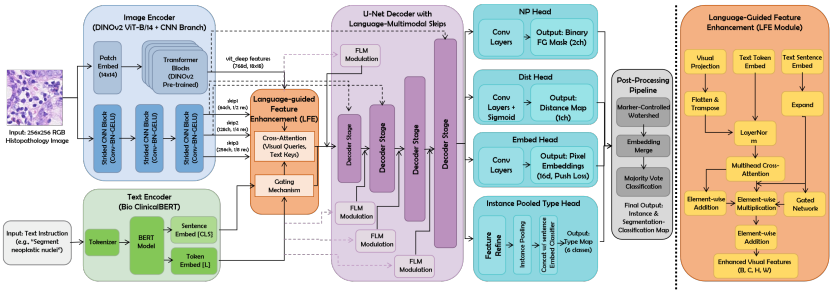

Fig. 2. A detailed diagram of the CIPS-Net architecture, where the Language-Guided Feature Enhancement (LFE) module is shown on the left.

2 Methodology

2.1 Network Architecture

CIPS-Net uses a dual-encoder single-decoder design. The visual encoder is DINOv2 ViT-B/14 [15], fine-tuned end-to-end. We extract activations from four equally-spaced layers and project them via convolutional adapters to obtain hierarchical features $\mathbf{F}_1$ to $\mathbf{F}_4$ for U-Net skip connections. Concurrently, a lightweight CNN branch captures fine-grained textural details missed by ViT's downsampled patches, aiding accurate boundary segmentation.

For text encoding, we use Bio_ClinicalBERT [2], keeping it frozen during training to leverage pre-trained clinical knowledge, prevent catastrophic forgetting, and reduce computational cost. Compared to one-hot vectors that encode predefined class IDs, natural language provides a broader semantic space that can capture clinical synonyms (e.g., "immune" vs. "inflammatory"). This is intended to allow pathologists to query using professional terminology and lay the groundwork for complex descriptive queries encompassing morphological, spatial, or functional characteristics. The text encoder produces per-token embeddings $\mathbf{E}_t \in \mathbb{R}^{L\times 512}$ for cross-modal attention and a sentence embedding $\mathbf{e}_s \in \mathbb{R}^{512}$ for skip connections.

The Language-Guided Feature Enhancement module (Fig. 2, orange) grounds instructions via multi-head cross-attention [22] at the encoder bottleneck. With $\mathbf{F}_4$ as queries and $\mathbf{E}_t$ as keys/values, each spatial position attends to specific tokens. This per-token attention aims to provide finer control to amplify relevant visual features compared to a global categorical switch: $\hat{\mathbf{F}}_4 = \mathbf{F}_4 + \mathrm{MHA}(\mathbf{F}_4, \mathbf{E}_t, \mathbf{E}_t)$. A sigmoid-activated gating mechanism, conditioned on $\mathbf{e}_s$, scales the cross-attention output before residual addition, ensuring text modulates rather than overrides visual features.

The decoder (Fig. 2, purple) follows a U-Net structure [17]. At each stage, the sentence embedding $\mathbf{e}_s$ generates per-channel scale and shift parameters via FiLM conditioning [16]. These dynamically re-weight encoder skip features, embedding the language instruction deeply into multi-scale spatial feature propagation rather than treating it as a post-hoc global filter.

2.2 Output Heads

Four lightweight convolutional heads operate on the shared decoder features (Fig. 2, red), each addressing a distinct aspect of instance segmentation. The nucleus presence head produces a two-class softmax indicating foreground versus background, supervised with a combination of focal loss and Dice loss to handle the severe class imbalance where nuclei occupy a small fraction of the image.

For instance separation, we regress a normalized Euclidean distance transform through a sigmoid-activated head, where each foreground pixel is assigned a value proportional to its distance from the nearest instance boundary, normalized per-instance so that centres achieve 1.0 and boundaries approach 0. Unlike HoVer-Net's horizontal-vertical maps [7], the local maxima of this single-channel representation directly provide seed markers for marker-controlled watershed segmentation without requiring Sobel gradient computation. The distance transform is supervised with mean squared error on foreground pixels together with a Sobel gradient penalty that sharpens predicted boundaries.

The instance embedding head predicts a 16-dimensional vector per pixel, trained with the discriminative loss [3] (Eq. 1):

$$\begin{aligned}\mathcal{L}_{\text{embed}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\Omega_k|}\sum_{p\in\Omega_k}\max(0,\|\hat{\mathbf{e}}_p-\boldsymbol{\mu}_k\|-\delta_v)^2 \\ +\frac{1}{K(K-1)}\sum_{\substack{k_a,k_b\\k_a\neq k_b}}\max(0,2\delta_d-\|\boldsymbol{\mu}_{k_a}-\boldsymbol{\mu}_{k_b}\|)^2\end{aligned} \tag{1}$$

where the pull term ($\delta_v = 0.5$) clusters same-instance pixels and the push term ($\delta_d = 1.5$) separates different instances. These heads interact synergistically: the distance transform provides spatial markers for watershed, while instance embeddings provide the semantic space to merge over-segmented fragments via union-find.

Finally, the type classification head pools decoder features within each ground truth instance region during training, concatenates the whole-nucleus representation with the text sentence embedding $\mathbf{e}_s$, and classifies through an MLP. This instance-pooled formulation captures holistic nuclear morphology rather than classifying pixels independently. An auxiliary pixel-level head provides dense supervision, and final types are assigned by per-instance majority voting. Weighted focal loss [12] addresses class imbalance. Fig. 2 presents an overview of CIPS-Net. The model takes a histopathology image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a natural language instruction specifying desired nucleus types, and outputs instance segmentation masks exclusively for the queried cell categories.

2.3 Permutation-Based Training

Since PanNuke [6] provides masks without text descriptions, we use MedGemma [19] to generate descriptions based on tissue type and nucleus classes, yielding 7,558 pairs (excluding 343 empty background patches). To teach conditional behaviour,

we enumerate all $2^N - 1$ class subsets for each image containing N classes, pairing each with a partial ground truth mask. An image with three classes yields seven combinations, expanding the training set to 30,828 pairs. Because nucleus types in histopathology are highly constrained (e.g., PanNuke has a maximum of 5 classes), the maximum possible subsets per image is only 31 (2^5-1), making this combinatorial strategy computationally efficient and perfectly tractable for this domain. This strategy encourages the model to segment mainly what the text specifies. Synthesizing queries from existing masks mitigates the need for large-scale text-image datasets, making this approach highly viable. For standardized quantitative benchmarking against unconditional "segment-all" baselines, evaluation is performed using a comprehensive full-class instruction, though the model retains its core flexibility for single-class selective querying in clinical practice (as demonstrated qualitatively).

2.4 Training and Inference

The total training loss combines the four head objectives[3] (Eq. 2):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{np}}\mathcal{L}_{\text{np}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{embed}}\mathcal{L}_{\text{embed}} + \lambda_{\text{type}}\mathcal{L}_{\text{type}} \tag{2}$$

Deferred re-weighting activates class-specific weights after 30% of training to prevent dominant classes from overwhelming early feature learning. At inference, the model receives the full-class instruction and produces four output maps from which instances are extracted by thresholding the NP map, detecting distance transform peaks as seed markers, applying marker-controlled watershed, merging adjacent instances via embedding-based union-find, filtering by size, and assigning types via per-instance majority voting. Merging occurs when adjacent regions' mean embeddings fall within the 1.25 threshold; reported FPS covers the full pipeline, including text encoding and post-processing.

3 Experiments and Results

Dataset and Setup. We evaluate on PanNuke [5, 6], comprising 7,901 H&E images (256×256) across 19 tissue types with 189,744 nuclei in five classes. We use the official 3-fold cross-validation protocol. Models are trained with AdamW (lr=10^{-4}, weight decay=10^{-4}), cosine annealing, batch size 8, and early stopping (patience 8) for 100 epochs on a single RTX A5000. Post-processing thresholds: NP=0.40, distance peak=0.60, embedding distance=1.25. We report multi-class panoptic quality (mPQ) as primary metric, plus binary PQ (bPQ) and AJI[4].

[3] Loss weights ($\lambda_{\text{np}} = 1, \lambda_{\text{dist}} = 2, \lambda_{\text{embed}} = 1, \lambda_{\text{type}} = 2$) were selected via a manual sweep over $\lambda \in \{0.5, 1, 2, 4\}$ on a held-out fold, chosen to balance gradient magnitudes across the multi-task objectives, ensuring the network sufficiently prioritizes the more challenging distance regression and typing tasks.

[4] We use the metrics code from the HoVer-Net GitHub: `https://github.com/vqdang/hover_net/tree/master/metrics`

Baselines. We compare against the following unconditional methods: HoVer-Net [7], StarDist [18], CPP-Net [4], PointNu-Net [26], CellViT [10], and Prompt-NucSeg [20]. To ensure a fair comparison for text-conditional VLM baselines (LAVT [25], CRIS [23], Grounding DINO [13], and LViT [11]), we replaced their native object predictors with standard HoVer-Net heads (nucleus presence and HV maps), operating directly on their fused multimodal decoder features, since native predictors target single-mask referring segmentation and are incompatible with PanNuke's multi-instance evaluation.

3.1 Comparison with State-of-the-Art

Table 1 compares CIPS-Net against VLM baselines. CIPS-Net achieves an mPQ of 0.4846 and a bPQ of 0.6217, outperforming LViT (0.4405) and Grounding DINO. In Table 2, CIPS-Net matches unconditional methods like CPP-Net (0.4847) while enabling text-conditional selective segmentation. Regarding efficiency, while CIPS-Net has 202.6 M total parameters, only 94.3 M are trainable. Crucially, CIPS-Net requires 11% fewer MACs (52.6 G vs. 59.0 G) and runs 20% faster than the top-performing PromptNucSeg-H, proving text-conditioning preserves inference efficiency. Fig. 3 illustrates CIPS-Net's consistent segmentation across Prostate, Lung, and Colon tissues.

3.2 Per-Class Analysis

CIPS-Net achieves competitive PQ values for the Inflammatory and Dead classes (0.471 and 0.211, respectively) compared to other methods, showing improvements over PromptNucSeg-H by 0.030 and 0.050 PQ points, respectively (see Table 3). This suggests that a conditional focus may aid in the discrimination of challenging, rare populations that "segment-all" approaches might not prioritize as effectively. This is clinically significant, as discriminating inflammatory cells is critical for evaluating the tumor immune microenvironment, while dead/necrotic cells are key indicators of tumor grade. We also provide the classification report in Table 3; CIPS-Net achieves a detection F1 score of 0.79, comparing favorably with VLM baselines (LViT: 0.72, LAVT: 0.67) and approaching unconditional methods. Per-class F1 scores are competitive with CPP-Net and HoVer-Net while generally exceeding the VLM baselines.

3.3 Ablation Studies

Table 4 presents component ablations (left) and backbone comparisons (right). A no-text baseline (mPQ=0.274) improves to 0.388 with static text conditioning, and permutation training raises it to 0.424 by forcing partial-class queries. Replacing HV-map heads with instance embeddings yields 0.481, and upgrading to DINOv2 ViT-B/14 achieves 0.485. DINOv2 outperforms supervised ViT-B/16, while permutation training provides consistent gains across all backbones.

Table 1. Tissue-wise binary Panoptic Quality (bPQ) and mean Panoptic Quality (mPQ) on the PanNuke dataset comparing text-conditional vision–language baselines (LAVT, CRIS, Grounding DINO, and LViT) against the proposed CIPS-Net.

Tissue	LAVT [25]		CRIS [23]		Grnd. DINO [13]		LViT [11]		**CIPS-Net (Ours)**	
	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ
Adrenal	0.5002	0.3887	0.4441	0.2936	0.5201	0.3994	0.6096	0.4535	**0.6549**	**0.5132**
Bile Duct	0.4390	0.3430	0.3937	0.2998	0.4838	0.3812	0.5748	0.4712	**0.6354**	**0.4986**
Bladder	0.4825	0.4347	0.4306	0.3980	0.5267	0.5095	0.6468	**0.6013**	**0.6790**	0.5746
Breast	0.4447	0.3635	0.3978	0.3182	0.4711	0.3827	0.5436	0.4404	**0.6266**	**0.5016**
Cervix	0.4545	0.3548	0.4184	0.3127	0.5009	0.3935	0.5899	0.4571	**0.6321**	**0.4925**
Colon	0.3227	0.2698	0.2901	0.2329	0.3576	0.2935	0.4447	0.3690	**0.5205**	**0.4252**
Esophagus	0.4299	0.3679	0.3342	0.2735	0.4276	0.3730	0.5266	0.4539	**0.6102**	**0.5182**
Head & Neck	0.3952	0.3358	0.3066	0.2420	0.4216	0.3448	0.5542	0.4593	**0.6043**	**0.4765**
Kidney	0.4190	0.3421	0.3809	0.2767	0.4875	0.3970	0.5844	0.4839	**0.6498**	**0.5264**
Liver	0.5222	0.4062	0.4690	0.3269	0.5652	0.4153	0.6404	0.4767	**0.6832**	**0.5142**
Lung	0.3450	0.2550	0.2845	0.1887	0.4042	0.2898	0.5136	0.3492	**0.5763**	**0.3850**
Ovarian	0.3710	0.3142	0.3003	0.2489	0.3852	0.3232	0.5072	0.4213	**0.6056**	**0.5093**
Pancreatic	0.4265	0.3330	0.3844	0.2901	0.4715	0.3707	0.5664	0.4500	**0.6225**	**0.5038**
Prostate	0.3767	0.3266	0.3127	0.2717	0.3877	0.3442	0.5093	0.4372	**0.6325**	**0.5328**
Skin	0.3200	0.2361	0.3090	0.1917	0.4246	0.3039	0.5330	0.3380	**0.5894**	**0.4146**
Stomach	0.3862	0.2582	0.3871	0.2465	0.4770	0.3318	0.5810	0.3952	**0.6472**	**0.4281**
Testis	0.4323	0.3270	0.3841	0.2802	0.4662	0.3596	0.5707	0.4278	**0.6456**	**0.5051**
Thyroid	0.4507	0.3143	0.4023	0.2750	0.4782	0.3521	0.5868	0.4349	**0.6618**	**0.4788**
Uterus	0.3246	0.2952	0.3041	0.2615	0.3991	0.3455	0.4897	**0.4496**	**0.5356**	0.4095
Average	0.4128	0.3298	0.3649	0.2752	0.4556	0.3637	0.5565	0.4405	**0.6217**	**0.4846**

Table 2. Comparison with unconditional SOTA on PanNuke (3-fold CV). †: text-conditional. **Bold**: best per column. Trainable parameters mentioned in brackets.

Method	Params (M)	MACs (G)	FPS	mPQ	bPQ
StarDist [18]	122.8	263.6	17	0.4625	0.6485
HoVer-Net [7]	37.6	150.0	7	0.4629	0.6596
CPP-Net [4]	122.8	264.4	14	0.4847	0.6798
CellViT-256 [10]	142.9	232.0	20	0.4850	0.6700
PointNu-Net [26]	158.1	335.1	11	0.4946	0.6810
PromptNucSeg-H [20]	145.6	59.0	27	**0.5123**	**0.6924**
CIPS-Net† (Ours)	202.6 (94.3)	52.6	32.3	0.4846	0.6217

Table 3. Detection F1 and per-class performance on PanNuke. Per-class entries are formatted as PQ (F1). **Bold**: best per column.

Method	Det-F1	Neo	Epi	Inf	Con	Dead
Mask-RCNN [9]	0.72	0.472(0.59)	0.403(0.52)	0.290(0.50)	0.300(0.42)	0.069(0.22)
StarDist [18]	0.82	0.547(0.69)	0.532(0.70)	0.424(0.57)	0.380(0.51)	0.123(0.10)
HoVer-Net [7]	0.80	0.551(0.62)	0.491(0.56)	0.417(0.54)	0.388(0.49)	0.139(0.31)
CPP-Net [4]	0.82	0.571(0.70)	0.565(0.72)	0.405(0.58)	0.395(0.53)	0.131(0.38)
CellViT-H [10]	0.83	0.581(**0.71**)	**0.583**(0.72)	0.417(0.58)	0.423(0.53)	0.149(0.36)
PromptNucSeg-H [20]	**0.84**	**0.598**(**0.71**)	0.582(**0.76**)	0.441(**0.59**)	**0.433**(**0.55**)	0.161(**0.46**)
LAVT [25]	0.67	0.358(0.53)	0.293(0.51)	0.332(0.50)	0.291(0.49)	0.068(0.34)
CRIS [23]	0.62	0.290(0.48)	0.239(0.45)	0.317(0.44)	0.240(0.43)	0.000(0.00)
LViT [11]	0.72	0.467(0.60)	0.435(0.61)	0.440(0.55)	0.356(0.50)	0.197(0.39)
CIPS-Net (Ours)	0.79	0.541(0.68)	0.539(0.70)	**0.471**(0.58)	0.397(0.54)	**0.211**(0.32)

Table 4. Ablation studies on PanNuke. Left: component ablation (HV: HoVer-Net heads; IE: instance embedding heads). Right: backbone comparison under static vs. permutation training. * denotes DINOv2-pretrained ViT (Base/Large).

Configuration	Text	Perm	Head	mPQ	bPQ	Backbone	Perm	mPQ	bPQ
No-text baseline	×	×	HV	0.274	0.378	ViT-B	×	0.388	0.583
+ Text cond. (static)	✓	×	HV	0.388	0.583	ViT-B	✓	0.481	0.608
+ Permutation training	✓	✓	HV	0.424	0.527	ViT-B*	×	0.475	0.600
+ Instance Embed. heads	✓	✓	IE	0.481	0.608	ViT-B*	✓	**0.485**	**0.622**
+ DINOv2 ViT-B*	✓	✓	IE	**0.485**	**0.622**	ViT-L*	✓	0.482	0.611

4 Conclusion

In this work, we introduced CIPS-Net, a text-instructed framework designed to perform conditional nucleus instance segmentation in histopathology images. By tightly integrating visual features from DINOv2 with rich semantic guidance from Bio_ClinicalBERT, CIPS-Net offers an intuitive alternative to traditional "segment-all" paradigms. This enables pathologists to extract clinically relevant cells on-demand, reducing cognitive load. Furthermore, our permutation-based training synthesizes class subsets from existing masks, teaching conditional behaviour without new manual labelling. Extensive evaluations on PanNuke show CIPS-Net trades a small mPQ gap to the top unconditional method (PromptNucSeg-H) for text-conditional selectivity and higher efficiency, while improving discrimination of rare inflammatory and dead cells. CIPS-Net's cross-attention/FiLM design is target-agnostic and may extend beyond nuclei to other instruction-driven pathology tasks, with potential integration into WSI viewers for clinical use. Our evaluation is limited to PanNuke; quantitative leakage across query subsets and robustness to instruction phrasing are left to future work.

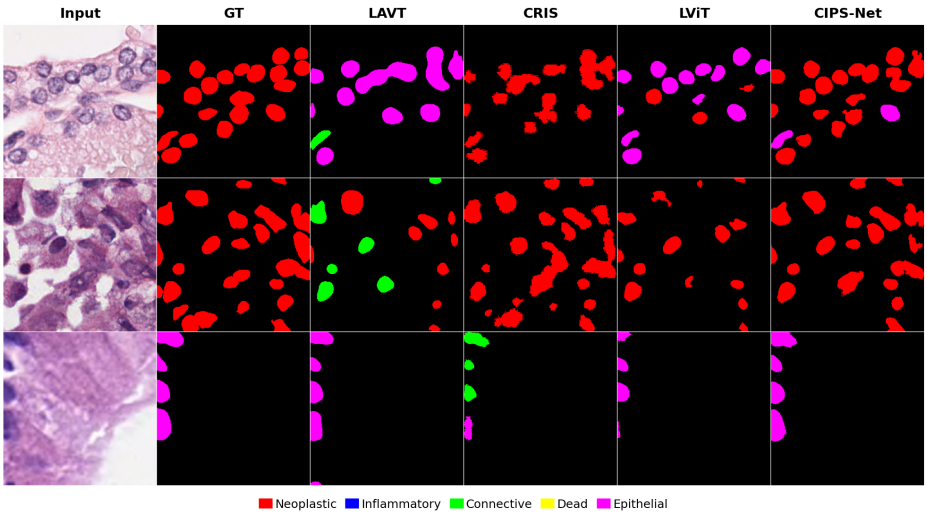

Fig. 3. Qualitative comparison. Columns: Input, GT, LAVT, CRIS, LViT, CIPS-Net.

References

1. Alrubyli, Y., Bevilacqua, A.: Cellvlm: Hierarchical text-guided vision transformers for cell segmentation and classification. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3432–3435 (2025). https://doi.org/10.1109/BIBM66473.2025.11356963
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/W19-1909, https://aclanthology.org/W19-1909/
3. Brabandere, B., Neven, D., Van Gool, L.: Semantic instance segmentation with a discriminative loss function (08 2017). https://doi.org/10.48550/arXiv.1708.02551
4. Chen, S., Ding, C., Liu, M., Cheng, J., Tao, D.: Cpp-net: Context-aware polygon proposal network for nucleus segmentation. IEEE Transactions on Image Processing **32**, 980–994 (2023). https://doi.org/10.1109/TIP.2023.3237013
5. Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A., Rajpoot, N.: Pannuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification. In: Digital Pathology. pp. 11–19. Springer International Publishing, Cham (2019)
6. Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N.: Pannuke dataset extension, insights and baselines (2020), https://arxiv.org/abs/2003.10778
7. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis **58**, 101563 (2019). https://doi.org/https://doi.org/10.1016/j.media.2019.101563, https://www.sciencedirect.com/science/article/pii/S1361841519301045

8. He, H., Huang, Z., Ding, Y., Song, G., Wang, L., Ren, Q., Wei, P., Gao, Z., Chen, J.: Cdnet: Centripetal direction network for nuclear instance segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4006–4015 (2021). https://doi.org/10.1109/ICCV48922.2021.00399
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
10. Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., Kleesiek, J.: Cellvit: Vision transformers for precise cell segmentation and classification. Medical Image Analysis **94**, 103143 (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103143, https://www.sciencedirect.com/science/article/pii/S1361841524000689
11. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: Language meets vision transformer in medical image segmentation (2023), https://arxiv.org/abs/2206.14718
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007 (2017). https://doi.org/10.1109/ICCV.2017.324
13. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino withnbsp;grounded pre-training fornbsp;open-set object detection. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII. p. 38–55. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72970-6_3, https://doi.org/10.1007/978-3-031-72970-6_3
14. Naylor, P., Laé, M., Reyal, F., Walter, T.: Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Transactions on Medical Imaging **38**(2), 448–459 (2019). https://doi.org/10.1109/TMI.2018.2865709
15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
16. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: visual reasoning with a general conditioning layer. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'18/IAAI'18/EAAI'18, AAAI Press (2018)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
18. Schmidt, U., Weigert, M., Broaddus, C., Myers, G.: Cell detection with star-convex polygons. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 265–273. Springer International Publishing, Cham (2018)
19. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
20. Shui, Z., Zhang, Y., Yao, K., Zhu, C., Zheng, S., Li, J., Li, H., Sun, Y., Guo, R., Yang, L.: Unleashing the power of prompt-driven nucleus instance segmentation. In: Computer Vision – ECCV 2024. pp. 288–304. Springer Nature Switzerland, Cham (2025)

21. Tang, J., Qian, B., Wan, P., Shao, W., Zhang, D.: LTSE: Language-guided Tissue Referring Segmentation in Pathology Images with Adaptive Expert Mixture . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15965. Springer Nature Switzerland (September 2025)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
23. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation (2022), https://arxiv.org/abs/2111.15174
24. Xing, F., Yang, L.: Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. IEEE Reviews in Biomedical Engineering **9**, 234–263 (2016). https://doi.org/10.1109/RBME.2016.2515127
25. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.S.: Lavt: Language-aware vision transformer for referring image segmentation (2022), https://arxiv.org/abs/2112.02244
26. Yao, K., Huang, K., Sun, J., Hussain, A.: Pointnu-net: Keypoint-assisted convolutional neural network for simultaneous multi-tissue histology nuclei segmentation and classification. IEEE Transactions on Emerging Topics in Computational Intelligence **8**(1), 802–813 (2024). https://doi.org/10.1109/TETCI.2023.3281864
27. Yellapragada, S., Graikos, A., Prasanna, P., Kurc, T., Saltz, J., Samaras, D.: Pathldm: Text conditioned latent diffusion model for histopathology. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5170–5179 (2024). https://doi.org/10.1109/WACV57701.2024.00510