UniNuc – Unified Automatic and Interactive Nucleus Instance Segmentation in Histopathology

Chao Qin[1] CHAO.QIN@MBZUAI.AC.AE
Fahad Shahbaz Khan[1] FAHAD.KHAN@MBZUAI.AC.AE
Jiale Cao[2] CONNOR@TJU.EDU.CN
Huazhu Fu[3] HZFU@IEEE.ORG
Shadab Khan[4] SHADAB.KHAN@ADIALAB.AE
Rao Muhammad Anwer[1] RAO.ANWER@MBZUAI.AC.AE
[1] *Department of Computer Vision, Mohamed bin Zayed Uni. of AI, United Arab Emirates*
[2] *School of Electrical and Information Engineering, Tianjin University, China*
[3] *Institute of High Performance Computing, A*STAR, Singapore*
[4] *Health Science, ADIA Lab, United Arab Emirates*

Editors: Under Review for MIDL 2026

Abstract

Accurate nucleus instance segmentation is a foundational task for computational pathology, yet dense cellularity, stain variability, and subtle boundaries make fully automatic pipelines prone to errors that must be efficiently corrected. Promptable segmentation foundation models such as the SAM provide an interface for human-in-the-loop refinement, but existing histopathology adaptations either prioritize fully automatic segmentation (often scaling to SAM-Huge for performance) or recover interactivity while retaining coarse decoding and bottom-up instance grouping that can fail in crowded tissue. We present **UniNuc**, a unified model that supports both automatic nucleus instance segmentation and iterative interactive refinement through a shared prompt interface. UniNuc (i) adopts an efficient SAM2 Hiera-B+ encoder together with a multi-scale high-quality mask decoder to preserve fine nuclear boundaries, and (ii) replaces heuristic pixel grouping with a DETR-style nuclei detector using a dedicated detection backbone whose predicted boxes serve as "auto-prompts". Optional language priors further improve nuclei type assignment. On PanNuke, UniNuc achieves 0.702 bPQ and 0.529 mPQ (0.548 mPQ with language priors), outperforming PromptNucSeg-H and CellViT-H while using substantially less compute than SAM-Huge-based baselines. On 14 wide-ranging datasets, UniNuc consistently improves interactive segmentation over PathoSAM-L in both in-domain and out-of-domain settings. Code and models will be publicly released.

Keywords: histopathology, instance segmentation, foundation model

1. Introduction

Nucleus instance segmentation is a foundational task in computational pathology, enabling quantitative tissue analysis for grading and tumor microenvironment profiling. Multi-organ benchmarks such as PanNuke and MoNuSAC have accelerated progress across diverse tissues and nucleus types (Gamper et al., 2019a; Verma et al., 2021), but dense clusters, variable staining, and ambiguous boundaries still cause frequent merge/split errors and hinder robust deployment. Promptable foundation models, most notably SAM (Kirillov

et al., 2023), offer an appealing pathway toward clinician-in-the-loop correction, motivating recent SAM adaptations for histopathology (Hörst et al., 2024; Shui et al., 2024).

However, existing approaches fall short of supporting pathology workflows in three main ways. **(S1) Human-in-the-loop refinement.** Automatic segmenters inevitably produce errors in complex tissue; without iterative interactive correction, clinicians lack the ability to refine segmentation (Alemi Koohbanani et al., 2020). Further, SAM-based methods usually target automatic instance segmentation and forgo iterative refinement (Hörst et al., 2024; Shui et al., 2024). **(S2) Crowded-tissue failure.** A large class of nuclei segmenters, including SAM-derived pipelines, still rely on bottom-up proxy-map regression (e.g., distance/offset/affinity maps) followed by watershed-style grouping (Naylor et al., 2019a; Graham et al., 2019; Chen et al., 2023; Yao et al., 2023; Griebel et al., 2025), which is brittle in crowded/overlapping nuclei. **(S3) Inefficient scaling and coarse decoding.** Prior SAM adaptations often scale to SAM-Huge for greater accuracy, incurring high cost with diminishing returns (Hörst et al., 2024; Shui et al., 2024); additionally, standard SAM-style decoders typically operate on coarse embeddings (e.g., $H/16$), which can blur small boundaries (Griebel et al., 2025; Ke et al., 2023). Lastly, recent evidence suggests detection and segmentation exploit different cues (corner/edge vs. semantics) (Li et al., 2025), complicating the use of a single backbone for both prompt generation and mask refinement.

To address these real-world gaps (S1)–(S3), we propose **UniNuc**, a unified prompt-based nuclei segmentation model that couples automatic instance segmentation with interactive refinement. UniNuc builds an interactive segmenter on the SAM2 Hiera-B+ encoder (Ravi et al., 2024) and introduces a *Multi-Scale High-Quality* mask decoder (Ke et al., 2023) that fuses $H/4$, $H/8$, and $H/16$ features via a learnt HQ token to recover precise boundaries. For automatic segmentation, UniNuc replaces heuristic grouping with a DETR-style nuclei detector (Chen et al., 2024; Robinson et al., 2025) and *decouples* feature extraction by using a dedicated detection backbone; predicted boxes are treated as "auto-prompts" and passed through the same prompt encoder and mask decoder used for user prompts. Optional language priors further improve nuclei type assignment. Experiments on PanNuke and 14 other datasets show that UniNuc achieves SOTA performance in both automatic/interactive settings while being computationally cheaper than SAM-L/Huge baselines.

2. Method

Fig. 1 summarizes **UniNuc**, which couples a SAM-style promptable segmenter with a nuclei detector that generates bounding boxes as *auto-prompts*. The segmentation core is shared across both modes: given an image x and a set of prompts $\mathcal{P}$ (points and/or boxes), UniNuc outputs a mask $\hat{m} = f(x, \mathcal{P})$. In interactive use, $\mathcal{P}$ is provided and refined by the user; in automatic use, $\mathcal{P}$ is produced by the detector. This shared interface ensures that the same prompt encoder and mask decoder are used for both workflows, and enables consistent refinement of automatic predictions.

2.1. Interactive Segmentation Core

The interactive segmenter follows SAM's decomposition into an image encoder, a prompt encoder, and a mask decoder (Kirillov et al., 2023). We adopt the **Hiera-B+** hierarchical ViT from SAM2 (Ravi et al., 2024) as the image encoder, which outputs a multi-scale

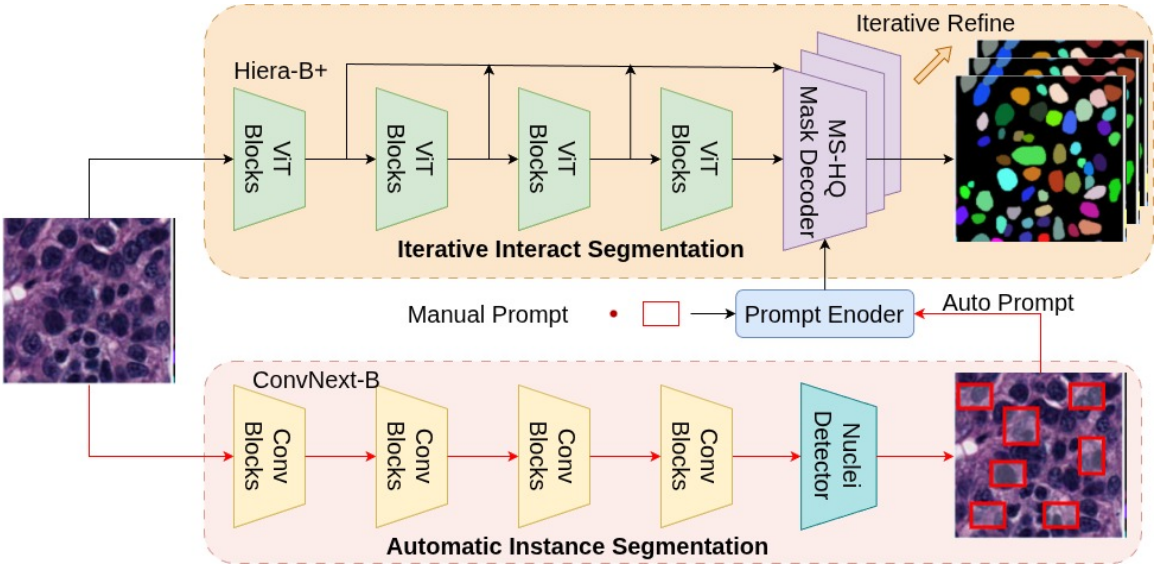

Figure 1: **UniNuc overview.** A single prompt-based segmentation core supports both (i) iterative interactive refinement from user prompts and (ii) automatic instance segmentation via *auto-prompts* produced by a nuclei detector. Optional language prompts can condition nuclei type predictions.

feature hierarchy. Let $\{F_0, F_1, F_2\}$ denote encoder features at resolutions $\{H/4, H/8, H/16\}$, respectively. The prompt encoder embeds a set of prompts $\mathcal{P} = \mathcal{P}_{\text{pt}} \cup \mathcal{P}_{\text{box}}$, where a point prompt is $p = (u, v, \ell)$ with label $\ell \in \{\text{fg}, \text{bg}\}$, and a box prompt is $b = (u_1, v_1, u_2, v_2)$ (normalized coordinates). These prompt embeddings are fused with image features by a high-quality multi-scale mask decoder (Sec. 2.2).

Iterative refinement simulation. To train robust interactive behavior, we simulate user corrections as in prior interactive segmentation work (Griebel et al., 2025). Given ground truth mask m and the current prediction $\hat{m}^{(t)}$, we compute false-negative and false-positive regions: $R_{\text{fn}}^{(t)} = m \setminus \hat{m}^{(t)}$ and $R_{\text{fp}}^{(t)} = \hat{m}^{(t)} \setminus m$. We sample a positive point from $R_{\text{fn}}^{(t)}$ (if non-empty) and a negative point from $R_{\text{fp}}^{(t)}$ (if non-empty), and update the prompt set: $\mathcal{P}^{(t+1)} = \mathcal{P}^{(t)} \cup \{p^+, p^-\}$. The refined mask is then $\hat{m}^{(t+1)} = f(x, \mathcal{P}^{(t+1)})$. We unroll this procedure for up to T iterations (we use T=8) during training, which encourages the decoder to incorporate corrective prompts effectively.

2.2. Multi-Scale High Quality Mask Decoder

Standard SAM-style decoders often operate on a single coarse feature map (typically $H/16$), which can blur small nuclei and boundary details in crowded tissue. UniNuc adopts a **Multi-Scale High Quality (MSHQ)** decoder inspired by SAM-HQ (Ke et al., 2023) (Fig. 2). Concretely, we introduce a learnable *HQ token* in addition to the standard SAM output token. Both tokens attend to the deep embedding F_2 via token–image cross-

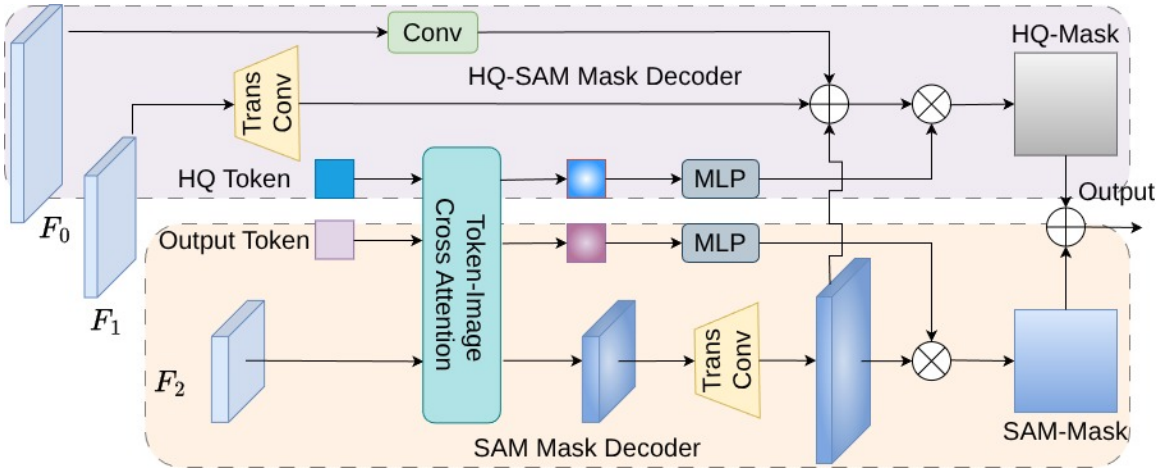

Figure 2: **Multi-Scale High Quality (MSHQ) mask decoder.** A learnable HQ token complements the standard output token and interacts with deep embeddings via token–image cross-attention. The updated deep features are fused with higher-resolution features (F_0, F_1) to recover boundary detail, and the HQ token produces the final high-quality mask.

attention, producing updated deep features $\tilde{F}_2$ and updated tokens. We then fuse $\tilde{F}_2$ with higher-resolution features via element-wise addition after lightweight projection to a common channel dimension:

$$\tilde{F} = \phi_0(F_0) \oplus \phi_1(F_1) \oplus \phi_2(\tilde{F}_2),$$

where $\phi_i(\cdot)$ are 1×1 projections (and upsampling where needed) and $\oplus$ is element-wise addition. The final high-quality mask is produced by projecting the HQ token with an MLP and combining it with $\tilde{F}$ (as in (Ke et al., 2023)), yielding sharper boundaries without requiring a larger backbone.

2.3. Task-Decoupled Dual-Backbone Design

Automatic instance segmentation requires both (i) accurate mask refinement (semantic, region-level cues) and (ii) reliable object localization to generate prompts (corner/edge cues). Recent analyses show that detection and segmentation emphasize different visual evidence and can compete when forced into a single shared representation (Li et al., 2025). UniNuc therefore **decouples** feature extraction: Hiera-B+ is optimized for prompt-conditioned mask refinement, while a dedicated ConvNeXt-B backbone extracts localization-friendly features for nuclei detection. This avoids optimizing one encoder for conflicting objectives and improves both the quality of auto-prompts and interactive refinement (see ablations).

2.4. Nuclei Detector for Automated Prompting

For automatic instance segmentation, UniNuc replaces heuristic bottom-up grouping with a **DETR-style nuclei detector** (Fig. 3) (Chen et al., 2024; Robinson et al., 2025). The

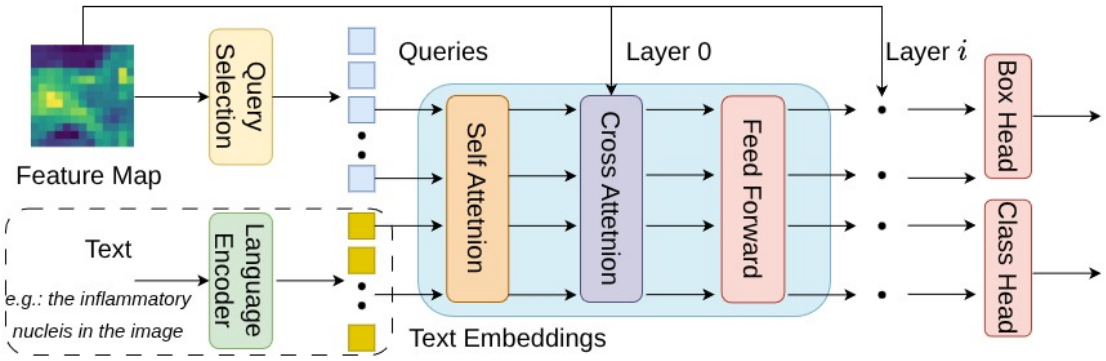

Figure 3: **Nuclei detector and auto-prompt generation.** A ConvNeXt-B backbone and DETR-style head produce box proposals. Top-K object queries are refined by a transformer decoder (optionally conditioned on text embeddings), predicting nucleus boxes and class logits. Predicted boxes are used as auto-prompts for the shared prompt encoder and mask decoder.

detector uses ConvNeXt-B features and a transformer decoder to produce a set of K object predictions. A lightweight scoring head first estimates a foreground score per location; we select the top-K locations to initialize object queries. The transformer decoder alternates self-attention among queries and cross-attention to image features, producing refined embeddings. Two heads then predict (i) bounding box b_k and (ii) class logits for each query.

Auto-prompts. Each predicted box b_k is converted into a SAM-style box prompt and passed to the shared prompt encoder and MSHQ decoder to yield an instance mask $\hat{m}_k$. Since DETR-style detectors already produce a fixed set of predictions without NMS, we retain the top-scoring boxes (or apply a confidence threshold) as the auto-prompt set.

Optional language priors. When nuclei categories are known (e.g., from dataset label sets or study context), we encode class names using a text encoder (e.g., CLIP text) and inject these embeddings into the detector's query self-attention by concatenation, allowing semantic priors to modulate query refinement and improve nuclei type assignment. The resulting class-aware boxes yield higher-quality auto-prompts, while mask generation remains prompt-driven and shared with interactive refinement.

3. Experiments and Results

3.1. Datasets and Protocols

We evaluate UniNuc on a benchmark of 14 public histopathology datasets. Following PathoSAM (Griebel et al., 2025), we train on 6 H&E datasets with instance annotations (Vu et al., 2019; Graham et al., 2021; Kumar et al., 2017; Gamper et al., 2019b; Schuiveling et al., 2025; Naylor et al., 2019b) and reserve 8 datasets strictly for out-of-domain testing (Graham et al., 2019; Mahbod et al., 2021; Sirinukunwattana et al., 2017; Naji et al., 2024; Verma et al., 2021; Alemi Koohbanani et al., 2020; Mahbod et al., 2024; Wang et al., 2024).

Table 1: Performance on the PanNuke dataset, using both binary PQ (bPQ) and multi-class PQ (mPQ) (Chen et al., 2023; Hörst et al., 2024). Best scores in **bold**.

Tissue	StarDist (Schmidt et al., 2018)		Hover-Net (Graham et al., 2019)		CPP-Net (Chen et al., 2023)		PointNu-Net (Yao et al., 2023)		CellViT-H (Hörst et al., 2024)		PromptNucSeg-H (Shui et al., 2024)		Ours	
	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ
Adrenal	0.6972	0.4868	0.6962	0.4812	0.7066	0.4944	0.7134	0.5115	0.7086	0.5134	0.7227	0.5128	**0.7410**	**0.5410**
Bile Duct	0.6690	0.4651	0.6696	0.4714	0.6768	0.4670	0.6814	0.4868	0.6784	0.4887	0.6976	0.5012	**0.7167**	**0.5250**
Bladder	0.6986	0.5793	0.7031	0.5792	0.7053	0.5936	0.7226	0.6065	0.7068	0.5844	0.7212	0.6043	**0.7271**	**0.6085**
Breast	0.6666	0.5064	0.6470	0.4902	0.6747	0.5090	0.6709	0.5147	0.6748	0.5180	**0.6842**	0.5322	0.6807	**0.5396**
Cervix	0.6690	0.4628	0.6652	0.4438	0.6912	0.4792	0.6899	0.5014	0.6872	0.4984	0.6983	0.5118	**0.7118**	**0.5227**
Colon	0.5779	0.4205	0.5575	0.4095	0.5911	0.4315	0.5945	0.4509	0.5921	0.4485	0.6096	0.4690	**0.6220**	**0.4889**
Esophagus	0.6655	0.5331	0.6427	0.5085	0.6797	0.5449	0.6766	0.5504	0.6682	0.5454	0.6920	**0.5711**	**0.6929**	0.5616
Head & Neck	0.6433	0.4768	0.6331	0.4530	0.6523	0.4706	0.6546	0.4838	0.6544	0.4913	0.6695	0.5104	**0.6888**	**0.5326**
Kidney	0.6998	0.4880	0.6836	0.4424	0.7067	0.5194	0.6912	0.5066	0.7092	0.5366	0.7115	**0.5786**	**0.7131**	0.5654
Liver	0.7231	0.5145	0.7248	0.4974	0.7312	0.5143	0.7314	0.5174	0.7322	0.5224	0.7372	0.5333	**0.7443**	**0.5675**
Lung	0.6362	0.4128	0.6302	0.4004	0.6386	0.4256	0.6352	0.4048	0.6426	0.4314	**0.6580**	0.4398	0.6530	**0.4464**
Ovarian	0.6668	0.5205	0.6309	0.4863	0.6830	0.5313	0.6863	0.5484	0.6722	0.5390	0.6856	0.5442	**0.6903**	**0.5619**
Pancreatic	0.6601	0.4585	0.6491	0.4600	0.6789	0.4706	0.6791	0.4804	0.6658	0.4719	0.6863	0.4974	**0.6937**	**0.5358**
Prostate	0.6748	0.5067	0.6615	0.5101	0.6927	0.5305	0.6854	0.5127	0.6821	0.5321	0.6983	0.5456	**0.7001**	**0.5573**
Skin	0.6289	0.3610	0.6234	0.3429	0.6209	0.3574	0.6494	0.4011	0.6565	0.4339	0.6613	0.4113	**0.7004**	**0.4547**
Stomach	0.6944	0.4477	0.6886	0.4726	0.7067	0.4582	0.7010	0.4517	0.7022	0.4705	0.7115	0.4559	**0.7156**	**0.4882**
Testis	0.6869	0.4942	0.6890	0.4754	0.7026	0.4931	0.7058	0.5334	0.6955	0.5127	0.7151	0.5474	**0.7249**	**0.5524**
Thyroid	0.6962	0.4300	0.6983	0.4315	0.7155	0.4392	0.7076	0.4508	0.7151	0.4519	0.7218	0.4721	**0.7280**	**0.4865**
Uterus	0.6599	0.4480	0.6393	0.4393	0.6615	0.4794	0.6634	0.4846	0.6625	0.4737	0.6743	0.4955	**0.6910**	**0.5148**
Average	0.6692	0.4744	0.6596	0.4629	0.6798	0.4847	0.6808	0.4957	0.6793	0.4980	0.6924	0.5123	**0.7019**	**0.5287**

We use two training protocols to ensure fair comparisons; (1) **Interactive segmentation (IS)** – train jointly on the 6 training datasets and compare to PathoSAM-L on in-domain and out-of-domain benchmarks; (2) **Automatic instance segmentation (AIS)** – train exclusively on PanNuke (Gamper et al., 2019a), consistent with SOTA AIS baselines such as CellViT-H (Hörst et al., 2024) and PromptNucSeg-H (Shui et al., 2024).

3.2. Implementation Details

We train UniNuc using AdamW optimizer (lr $5{\times}10^{-5}$, weight decay 10^{-4}) for 40 epochs with cosine annealing lr scheduler. We adopt the CellViT augmentation pipeline (Hörst et al., 2024) (random flipping, cropping, resizing, and color jitter). For interactive segmentation, we supervise with Dice, Focal, and MSE losses:

$$\mathcal{L}_{IS} = \mathcal{L}_{dice} + \lambda_{focal}\mathcal{L}_{focal} + \mathcal{L}_{mse}, \tag{1}$$

and for the nuclei detector we use a DETR-style objective; a combination of classification loss (IA-BCE), Generalized IoU loss, and L1 regression loss:

$$\mathcal{L}_{Box} = \mathcal{L}_{cls} + \lambda_{iou}\mathcal{L}_{iou} + \lambda_{l1}\mathcal{L}_{l1}, \tag{2}$$

with λ_{focal}=20, λ_{iou}=5, and λ_{l1}=2.

3.3. Automatic Instance Segmentation on PanNuke

We evaluate AIS on PanNuke using Panoptic Quality (PQ), reporting binary PQ (bPQ) and multi-class PQ (mPQ) following established protocols (Gamper et al., 2019a; Chen et al., 2023; Hörst et al., 2024). Results are summarized in Table 1. UniNuc achieves **0.7019 bPQ** and **0.5287 mPQ**, outperforming PromptNucSeg-H (Shui et al., 2024) (0.6924 bPQ, 0.5123 mPQ) and CellViT-H (Hörst et al., 2024) (0.6793 bPQ, 0.4980 mPQ). Importantly, these gains are obtained *without* stain normalization, test-time augmentation, oversampling, or

Table 2: Comparison of interactive segmentation performance (mSA). We report results for initial prompts (I_0) and after 7 refinement iterations (I_7) for both point (p) and box (b) inputs. The top three datasets are in-domain (seen during training), while the bottom three are out-of-domain (unseen). The best mSA scores are highlighted in **bold**.

Model	Dataset		PanNuke				MoNuSeg				CPM17			
	Parameters(M)	FLOPS(G)	I_0^p	I_7^p	I_0^b	I_7^b	I_0^p	I_7^p	I_0^b	I_7^b	I_0^p	I_7^p	I_0^b	I_7^b
Sam-L(no train)	304	1312	0.1619	0.7388	0.6093	0.8101	0.2764	0.6194	0.5204	0.6634	0.2482	0.7285	0.624	0.7689
PathoSam-L	304	1312	0.5399	0.9167	0.8057	0.9508	0.4595	0.8464	0.7049	0.8919	0.555	0.8342	0.7159	0.8783
Ours	69	264	**0.5668**	**0.9838**	**0.8626**	**0.9925**	**0.4683**	**0.9017**	**0.7289**	**0.9407**	**0.5601**	**0.9121**	**0.7656**	**0.9357**
Model	Dataset		LyNSeC(H&E)				LyNSeC(IHC)				NuClick			
	Parameters(M)	FLOPS(G)	I_0^p	I_7^p	I_0^b	I_7^b	I_0^p	I_7^p	I_0^b	I_7^b	I_0^p	I_7^p	I_0^b	I_7^b
Sam-L(no train)	304	1312	0.362	0.7126	0.6346	0.7392	0.3231	0.6544	0.5729	0.6958	0.2421	0.7555	0.6178	0.8097
PathoSam-L	304	1312	0.5765	0.9008	0.7744	0.9257	0.4399	0.8762	0.73	0.9131	0.3037	0.8621	0.7445	0.9216
Ours	69	264	**0.6169**	**0.9359**	**0.8147**	**0.9724**	**0.5511**	**0.9416**	**0.7772**	**0.9676**	**0.4267**	**0.9510**	**0.8252**	**0.9775**

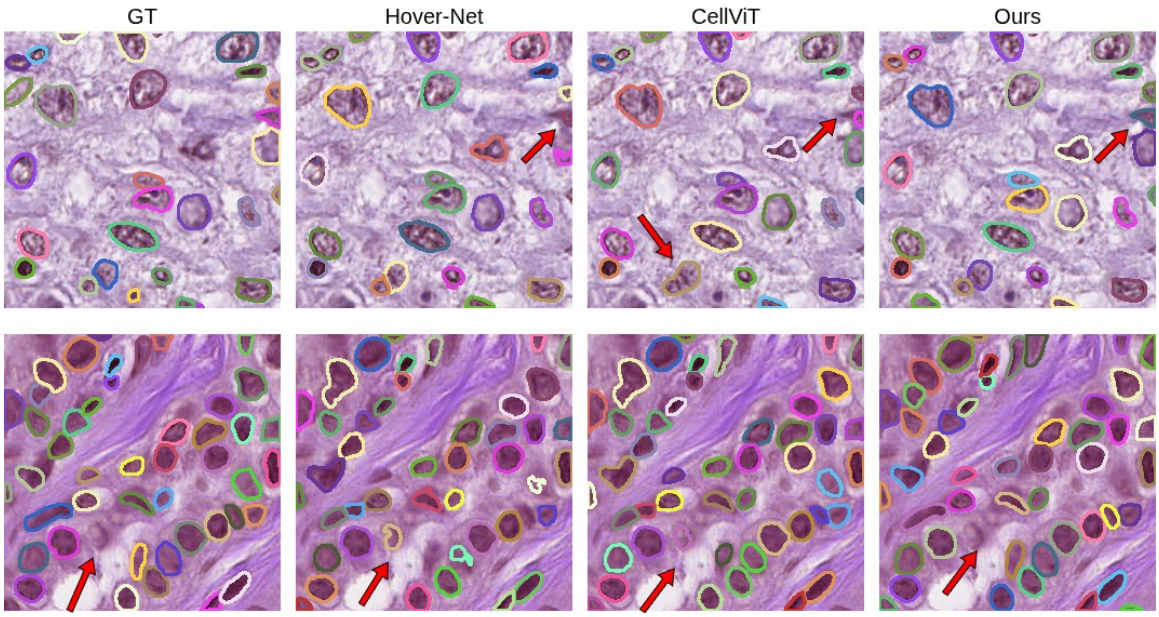

Figure 4: **Qualitative automatic instance segmentation (AIS) on PanNuke.** Compared to HoVer-Net (Graham et al., 2019) and CellViT-H (Hörst et al., 2024), UniNuc better separates crowded nuclei and preserves fine boundaries while maintaining correct nucleus categories. Best viewed zoomed in.

auxiliary tissue heads. When enabling language priors (Appendix A), mPQ further increases to **0.5484** while bPQ remains essentially unchanged (0.7022), indicating that language primarily improves nuclei type assignment rather than geometric mask quality.

Our design targets the inefficiency of brute-force scaling in histopathology. Although SAM/SAM2 provide strong initialization (Kirillov et al., 2023; Ravi et al., 2024), moving to SAM-Huge yields diminishing returns: PromptNucSeg-H and CellViT-H reach mPQ 0.5123/0.4980 versus 0.5095/0.4923 with SAM-Base, i.e., < 0.006 gain for a ~7× parameter increase (86M→632M) (Hörst et al., 2024; Shui et al., 2024). UniNuc instead uses SAM2's

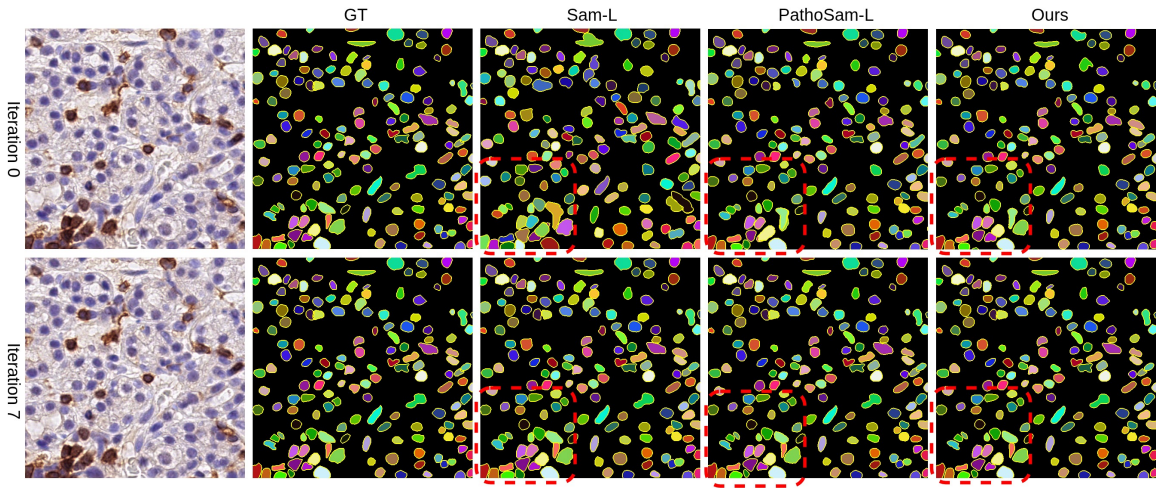

Figure 5: **Qualitative interactive segmentation on LyNSeC(IHC).** We compare SAM-L (Kirillov et al., 2023), PathoSAM-L (Griebel et al., 2025), and UniNuc at the initial prompt and after iterative refinement. UniNuc produces higher-quality initial masks and converges faster with corrective prompts.

efficient Hiera-B+ encoder (Ravi et al., 2024) and invests capacity in task-specific decoding and detection, improving the accuracy–efficiency trade-off.

Qualitative results (Fig.4) show that UniNuc can separate nuclei in crowded regions and preserves fine boundaries better than prior SOTA (HoVerNet, CellViT), consistent with the quantitative gains. We also report the metrics of our model with language priors in Appendix A.1.

3.4. Iterative Interactive Segmentation

We compare IS against SAM-L (no training) (Kirillov et al., 2023) and PathoSAM-L (Griebel et al., 2025) using mean Segmentation Accuracy (mSA) (Sec. B). Table 2 reports results at the initial prompt (I_0) and after 7 refinements (I_7) for both point (p) and box (b) prompts, across in-domain (PanNuke, MoNuSeg, CPM17) and out-of-domain datasets not seen during the training (LyNSeC-H&E, LyNSeC-IHC, NuClick). UniNuc consistently improves over PathoSAM-L across all settings while being substantially more efficient (69M params / 264 GFLOPs vs. 304M / 1312 GFLOPs).

The out-of-domain setting is most indicative of clinical transfer. On NuClick, UniNuc improves initial point accuracy from 0.3037 to **0.4267** and initial box accuracy from 0.7445 to **0.8252**. Similarly, on LyNSeC-H&E and LyNSeC-IHC, UniNuc improves initial point mSA from 0.5765 to 0.6169 and 0.4399 to 0.5511, and initial box mSA from 0.7744 to 0.8147 and 0.73 to 0.7772 (Table 4), which highlights UniNuc's robustness to staining and acquisition shifts. For AIS, language priors raise mPQ from 0.5287 to 0.5484 while keeping bPQ essentially unchanged (0.7019 to 0.7022) (Table 7). The full dual-backbone model uses 156M parameters / 584 GFLOPs (Table 4). Gains persist after refinement (e.g., I_7^p: 0.8621→**0.9510**). Fig. 5 illustrates that UniNuc produces better initial masks and requires fewer corrective prompts to approach the final refined quality.

Table 3: **Component Effectiveness Analysis.** Ablation study on the PanNuke dataset evaluating the Dual-Branch Encoder and Multi-Scale Mask Decoder. The baseline utilizes a single Hiera-B+ backbone and the original SAM mask decoder.

Components		Metrics – IS (mSA) / AIS (mSA)		
Dual-Branch	**Multi-Scale Mask Decoder**	I_0^p	I_7^p	AIS
-	-	0.5077	0.9348	0.4546
✓	-	0.5383	0.9631	0.4682
✓	✓	**0.5480**	**0.9730**	**0.4760**

Table 4: **Ablation of Backbone Size.** Comparison of model complexity and performance on PanNuke. Scaling up to the Large variant yields diminishing returns; thus, we select the Base+ combination as the optimal trade-off.

Backbone	Params(M)	FLOPS(G)	I_0^p	I_7^p	AIS
Hiera-S + ConvNext-S	83	316	0.5468	0.9697	0.4757
Hiera-B+ + ConvNext-B	156	584	0.5480	0.973	0.476
Hiera-L + ConvNext-L	409	1528	0.5505	0.9674	0.474

3.5. Prompt Efficiency

Interactive correction is only practical if a usable mask appears after a handful of clicks. Compared with PathoSAM-L, UniNuc starts closer to the target – on NuClick, it improves initial accuracy from I_0^p = 0.3037 to 0.4267 and from I_0^b = 0.7445 to 0.8252 (Table 2). Using the full refinement curves (Appendix A.2) we find that, UniNuc reaches PathoSAM-L's 7-click quality with substantially fewer corrections—typically 3–5 for box prompts (PanNuke: I_3^b = 0.9654 vs. PathoSAM I_7^b = 0.9508; NuClick: I_3^b = 0.9260 vs. PathoSAM I_7^b = 0.9216) and similarly reduces point refinements (PanNuke: I_4^p = 0.9388 vs. PathoSAM I_7^b = 0.9167). This faster convergence, coupled with lower compute (69M/264 GFLOPs vs. 304M/1312 GFLOPs), makes human-in-the-loop edits more feasible at scale.

3.6. Ablation Study

Table 3 evaluates the effect of UniNuc components on PanNuke, reporting I_0^p, I_7^p, and AIS performance (mSA). Starting from a baseline with a single Hiera-B+ encoder and the standard SAM decoder, adding the **dual-branch design** improves both IS (I_0^p: 0.5077→0.5383) and AIS (0.4546→0.4682), supporting the hypothesis that detection and segmentation benefit from different feature cues (Li et al., 2025). Adding the **multi-scale HQ decoder** yields further gains (I_0^p=0.5480, I_7^p=0.9730, AIS=0.4760), consistent with the need to preserve high-frequency boundary detail for small, crowded nuclei.

Table 4 studies three different backbone scaling in the dual-branch setting. The Large variant increases complexity sharply (409M params / 1528 GFLOPs) without improving

AIS and slightly degrades I_7^p, whereas the Base+ configuration provides the best overall trade-off. This phenomenon of performance saturation is consistent with findings reported in PathoSAM (Griebel et al., 2025). Consequently, we selected the Hiera-B+ /ConvNeXt-B combination as the default, as it offers the best balance between accuracy and efficiency.

Finally, we evaluate the **box-prompt strategy** for AIS. Most existing histopathology instance segmentation models regress proxy maps and rely on bottom-up grouping with post-processing, e.g., HoVer-Net predicts horizontal/vertical offsets (Graham et al., 2019) and PathoSAM predicts distance-to-center/boundary maps with watershed grouping (Griebel et al., 2025). In contrast, UniNuc uses a nuclei detector to generate box prompts in a top-down manner (Chen et al., 2024; Robinson et al., 2025). Replacing the detector with a bottom-up U-Net distance-map reduces AIS (mSA 0.4760→0.4647), confirming that detection-led prompt generation is more robust than heuristic grouping in overlapping area.

4. Discussion and Conclusion

UniNuc is designed around clinically motivated constraints: automatic nucleus segmentation is never perfectly reliable in real tissue, and practical workflows require a shared interface for *both* high-throughput automatic processing and targeted interactive correction. Existing SAM adaptations tend to optimize one side of this trade-off. UniNuc addresses both sides and effectively fills the gaps by (i) retaining a promptable segmentation core and training it explicitly for iterative refinement, (ii) improving boundary fidelity through a multi-scale HQ decoder (Ke et al., 2023), and (iii) replacing bottom-up grouping with a detector that produces box *auto-prompts* that can be refined identically to user prompts. This yields consistent gains across both tasks: higher PQ on PanNuke (0.5287 mPQ; 0.5484 with language priors) and improved interactive mSA across in-domain and out-of-domain datasets, with substantially lower compute than SAM-L/Huge baselines.

A key empirical lesson is that *scaling alone is insufficient*: prior work reports < 0.006 mPQ gains when moving from SAM-Base to SAM-Huge (Hörst et al., 2024; Shui et al., 2024). UniNuc instead uses efficient pretrained components (Hiera-B+ from SAM2 (Ravi et al., 2024)) and allocates model capacity where it directly affects pathology performance: multi-scale decoding for fine boundaries and task-decoupled detection for reliable prompt generation (Li et al., 2025). Optional language priors further improve nuclei type assignment without changing the underlying mask geometry.

In terms of its limitations, UniNuc currently assumes reliable box proposals; missed detections can limit AIS, and language priors depend on the availability of meaningful class names. Future work should evaluate end-to-end WSI pipelines, incorporate uncertainty-aware prompting to prioritize ambiguous regions, and validate usability via prospective annotation studies with pathologists. In summary, UniNuc provides a practical unification of automatic instance segmentation and iterative interactive refinement for histopathology, improving both accuracy and efficiency while directly targeting failure modes in dense tissue.

References

Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: A deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, October 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101771. URL http://doi.org/10.1016/j.media.2020.101771.

Qiang Chen, Xiangbo Su, Xinyu Zhang, Jian Wang, Jiahui Chen, Yunpeng Shen, Chuchu Han, Ziliang Chen, Weixiang Xu, Fanrong Li, et al. Lw-detr: A transformer replacement to yolo for real-time detection. *arXiv preprint arXiv:2406.03459*, 2024.

Shengcong Chen, Changxing Ding, Minfeng Liu, Jun Cheng, and Dacheng Tao. Cpp-net: Context-aware polygon proposal network for nucleus segmentation. *IEEE Transactions on Image Processing*, 32:980–994, 2023.

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 *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer, 2019a.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, pages 11–19. Springer, 2019b. URL https://doi.org/10.1007/978-3-030-23937-4_2.

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.

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 (ICCV) Workshops*, pages 684–693, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021W/CDPath/html/Graham_Lizard_A_Large-Scale_Dataset_for_Colonic_Nuclear_Instance_Segmentation_and_ICCVW_2021_paper.html.

Titus Griebel, Anwai Archit, and Constantin Pape. Segment anything for histopathology. *arXiv preprint arXiv:2502.00408*, 2025.

Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94: 103143, 2024.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

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 Transactions on Medical Imaging*, 36(7):1550–1560, July 2017. ISSN 1558-254X. doi: 10.1109/tmi.2017.2677499. URL http://doi.org/10.1109/tmi.2017.2677499.

Yifan Li, Xin Li, Tianqin Li, Wenbin He, Yu Kong, and Liu Ren. Vit-split: Unleashing the power of vision foundation models via efficient splitting heads. *arXiv preprint arXiv:2506.03433*, 2025.

Amirreza Mahbod, Gerald Schaefer, Benjamin Bancher, Christine Löw, Georg Dorffner, Rupert Ecker, and Isabella Ellinger. Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. *Computers in Biology and Medicine*, 132:104349, May 2021. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2021.104349. URL http://doi.org/10.1016/j.compbiomed.2021.104349.

Amirreza Mahbod, Christine Polak, Katharina Feldmann, Rumsha Khan, Katharina Gelles, Georg Dorffner, Ramona Woitek, Sepideh Hatamikia, and Isabella Ellinger. Nuinsseg: A fully annotated dataset for nuclei instance segmentation in h&e-stained histological images. *Scientific Data*, 11(1), March 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03117-2. URL http://doi.org/10.1038/s41597-024-03117-2.

Hussein Naji, Lucas Sancere, Adrian Simon, Reinhard Büttner, Marie-Lisa Eich, Philipp Lohneis, and Katarzyna Bożek. Holy-net: Segmentation of histological images of diffuse large b-cell lymphoma. *Computers in Biology and Medicine*, 170:107978, March 2024. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2024.107978. URL http://doi.org/10.1016/j.compbiomed.2024.107978.

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, 2019a. doi: 10.1109/TMI.2018.2865709.

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, February 2019b. ISSN 1558-254X. doi: 10.1109/tmi.2018.2865709. URL http://doi.org/10.1109/tmi.2018.2865709.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. Rf-detr: Neural architecture search for real-time detection transformers. *arXiv preprint arXiv:2511.09554*, 2025.

Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 265–273. Springer, 2018.

Mark Schuiveling, Hong Liu, Daniel Eek, Gerben E Breimer, Karijn P M Suijkerbuijk, Willeke A M Blokx, and Mitko Veta. A novel dataset for nuclei and tissue segmentation in melanoma with baseline nuclei segmentation and tissue segmentation benchmarks. *GigaScience*, 14, 2025. ISSN 2047-217X. doi: 10.1093/gigascience/giaf011. URL http://doi.org/10.1093/gigascience/giaf011.

Zhongyi Shui, Yunlong Zhang, Kai Yao, Chenglu Zhu, Sunyi Zheng, Jingxiong Li, Honglin Li, Yuxuan Sun, Ruizhe Guo, and Lin Yang. Unleashing the power of prompt-driven nucleus instance segmentation. In *European conference on computer vision*, pages 288–304. Springer, 2024.

Korsuk Sirinukunwattana, Josien P.W. Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J. Matuszewski, Elia Bruni, Urko Sanchez, Anton Böhm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David R.J. Snead, and Nasir M. Rajpoot. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35:489–502, January 2017. ISSN 1361-8415. doi: 10.1016/j.media.2016.08.008. URL http://doi.org/10.1016/j.media.2016.08.008.

Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E. Ahmed Raza, Nasir Rajpoot, Xiyi Wu, Huai Chen, Yijie Huang, Lisheng Wang, Hyun Jung, G. Thomas Brown, Yanling Liu, Shuolin Liu, Seyed Alireza Fatemi Jahromi, Ali Asghar Khani, Ehsan Montahaei, Mahdieh Soleymani Baghshah, Hamid Behroozi, Pavel Semkin, Alexandr Rassadin, Prasad Dutande, Romil Lodaya, Ujjwal Baid, Bhakti Baheti, Sanjay Talbar, Amirreza Mahbod, Rupert Ecker, Isabella Ellinger, Zhipeng Luo, Bin Dong, Zhengyu Xu, Yuehan Yao, Shuai Lv, Ming Feng, Kele Xu, Hasib Zunair, Abdessamad Ben Hamza, Steven Smiley, Tang-Kai Yin, Qi-Rui Fang, Shikhar Srivastava, Dwarikanath Mahapatra, Lubomira Trnavska, Hanyun Zhang, Priya Lakshmi Narayanan, Justin Law, Yinyin Yuan, Abhiroop Tejomay, Aditya Mitkari, Dinesh Koka, Vikas Ramachandra, Lata Kini, and Amit Sethi. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12):3413–3423, December 2021. ISSN 1558-254X. doi: 10.1109/tmi.2021.3085712. URL http://doi.org/10.1109/TMI.2021.3085712.

Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, Rajarsi Gupta, Jin Tae Kwak, Nasir Rajpoot, Joel Saltz, and Keyvan Farahani. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in Bioengineering and Biotechnology*, 7, April 2019. ISSN 2296-4185. doi: 10.3389/fbioe.2019.00053. URL http://doi.org/10.3389/fbioe.2019.00053.

Ranran Wang, Yusong Qiu, Xinyu Hao, Shan Jin, Junxiu Gao, Heng Qi, Qi Xu, Yong Zhang, and Hongming Xu. Simultaneously segmenting and classifying cell nuclei by using multi-task learning in multiplex immunohistochemical tissue microarray sections. *Biomedical Signal Processing and Control*, 93:106143, July 2024. ISSN 1746-8094. doi: 10.1016/j.bspc.2024.106143. URL http://doi.org/10.1016/j.bspc.2024.106143.

Kai Yao, Kaizhu Huang, Jie Sun, and Amir Hussain. Pointnu-net: Keypoint-assisted convolutional neural network for simultaneous multi-tissue histology nuclei segmentation and classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

Table 5: **Iterative Performance with Point Prompts Inputs.** mSA scores across all 14 datasets for iterations I_0^p through I_7^p.

Dataset	I_0^p	I_1^p	I_2^p	I_3^p	I_4^p	I_5^p	I_6^p	I_7^p
CPM15	0.5443	0.6797	0.7529	0.8022	0.8377	0.8671	0.8929	0.9146
CPM17	0.5601	0.6783	0.7554	0.8044	0.8405	0.8720	0.8945	0.9121
CoNSeP	0.3257	0.4705	0.5998	0.7013	0.7742	0.8296	0.8674	0.8959
CryoNuSeg	0.3670	0.4885	0.5995	0.6991	0.7658	0.8185	0.8554	0.8876
Lizard	0.5090	0.6783	0.7927	0.8697	0.9203	0.9522	0.9713	0.9827
LyNSec(H&E)	0.6169	0.7354	0.8073	0.8541	0.8891	0.9153	0.9359	0.9521
LyNSec(IHC)	0.5511	0.6894	0.7726	0.8278	0.8677	0.8980	0.9222	0.9416
MoNuSAC	0.4232	0.5921	0.7028	0.7821	0.8388	0.8746	0.9012	0.9225
MoNuSeg	0.4683	0.6063	0.6975	0.7642	0.8146	0.8542	0.8816	0.9017
NuClick	0.4267	0.6231	0.7432	0.8223	0.8738	0.9076	0.9323	0.9510
NuInsSeg	0.3252	0.4641	0.5828	0.6799	0.7489	0.8017	0.8400	0.8695
PanNuke	0.5668	0.7351	0.8404	0.9014	0.9388	0.9613	0.9755	0.9838
Puma	0.5662	0.7061	0.7941	0.8535	0.8941	0.9237	0.9445	0.9605
TNBC	0.5707	0.6834	0.7555	0.8122	0.8566	0.8895	0.9093	0.9272

Appendix A. Detailed Results

A.1. AIS Performance with Language Priors

As previously described, our nuclei detector is designed to integrate semantic class priors. In clinical scenarios, the specific tissue type or potential nuclei classes are often known beforehand. To leverage this, we utilize a CLIP-text (Radford et al., 2021) encoder to map class names into language embeddings, which are then concatenated with the object queries. Table 7 presents the Automatic Instance Segmentation (AIS) results on the PanNuke dataset. Incorporating the language prior yields a substantial improvement in Multi-class Panoptic Quality (mPQ), raising the average score from 0.5287 to 0.5484. Notably, the Binary Panoptic Quality (bPQ) remains stable (0.7019 vs. 0.7022). This indicates that while the geometric quality of the segmentation masks remains consistent, the language prior significantly enhances the model's ability to correctly classify nuclei types across diverse tissues.

A.2. Extended Iterative Segmentation Results

In this section, we report the complete metrics for all iterative steps (iterations 0 through 7) across the test split of 14 datasets, which are presented in Table 5 and Table 6 to facilitate future benchmarking and comparison.

A.3. Qualitative Examples of Iterative Interactive Segmentation

We provide qualitative visualizaton of the interactive segmentation across all 14 datasets, as shown in Fig. 6 and Fig. 7.

Table 6: **Iterative Performance with Box Prompts Inputs.** mSA scores across all 14 datasets for iterations I_0^p through I_7^p.

Dataset	I_0^b	I_1^b	I_2^b	I_3^b	I_4^b	I_5^b	I_6^b	I_7^b
CPM15	0.7674	0.8081	0.8419	0.8676	0.8912	0.9088	0.9238	0.9363
CPM17	0.7656	0.8118	0.8437	0.8700	0.8918	0.9098	0.9244	0.9357
CoNSeP	0.6760	0.7573	0.8116	0.8522	0.8848	0.9062	0.9277	0.9409
CryoNuSeg	0.6600	0.7371	0.7956	0.8396	0.8768	0.9012	0.9223	0.9394
Lizard	0.7912	0.8647	0.9142	0.9486	0.9684	0.9799	0.9871	0.9919
LyNSec(H&E)	0.8147	0.8576	0.8886	0.9134	0.9335	0.9500	0.9629	0.9724
LyNSec(IHC)	0.7772	0.8295	0.8675	0.8968	0.9199	0.9391	0.9551	0.9676
MoNuSAC	0.7482	0.8070	0.8480	0.8833	0.9069	0.9246	0.9440	0.9584
MoNuSeg	0.7290	0.7860	0.8311	0.8646	0.8888	0.9110	0.9274	0.9407
NuClick	0.8252	0.8689	0.9011	0.9260	0.9468	0.9616	0.9702	0.9775
NuInsSeg	0.6557	0.7330	0.7887	0.8298	0.8615	0.8883	0.9074	0.9233
PanNuke	0.8626	0.9130	0.9450	0.9654	0.9776	0.9848	0.9897	0.9925
Puma	0.8212	0.8715	0.9057	0.9319	0.9499	0.9635	0.9738	0.9803
TNBC	0.7825	0.8271	0.8681	0.8926	0.9160	0.9323	0.9455	0.9585

A.4. AIS performance across 14 Datasets

We evaluate the automatic instance segmentation performance of UniNuc across the test splits of all 14 datasets. The comprehensive results are reported in Table 8 to serve as a baseline for future benchmarking and comparison.

Appendix B. Evaluation Metrics

In this section, we provide detailed definitions of the metrics used to evaluate both interactive and automatic segmentation performance.

B.1. Iterative Interactive Segmentation Metric

Following PathoSAM (Griebel et al., 2025), we utilize the Mean Segmentation Accuracy (mSA) to evaluate iterative interactive segmentation results. This metric relies on the count of True Positives (TP), False Negatives (FN), and False Positives (FP), which are derived from the Intersection over Union (IoU) between predicted and ground-truth objects. Specifically, at a given threshold t, a predicted object is considered a "match" (True Positive) if its IoU with a ground-truth object exceeds t. Consequently:

- $TP(t)$: The number of correctly matched objects ($IoU > t$).
- $FP(t)$: The number of predicted objects with no matching ground truth.
- $FN(t)$: The number of ground-truth objects with no matching prediction.

Table 7: **Effect of Language Priors on AIS Performance.** Comparison of Panoptic Quality scores on the PanNuke dataset. The inclusion of language priors significantly boosts the multi-class metric (mPQ), demonstrating improved nuclei classification accuracy.

	Metric	Adr.	Bile	Blad.	Bre.	Cerv.	Col.	Eso.	Head	Kid.	Liv.	Lung	Ova.	Pan.	Pros.	Skin	Stom.	Test.	Thy.	Ute.	Avg.
Default	bPQ	0.741	0.717	0.727	0.681	0.712	0.622	0.693	0.689	0.713	0.744	0.653	0.690	0.694	0.700	0.700	0.716	0.725	0.728	0.691	0.7019
	mPQ	0.541	0.525	0.609	0.540	0.523	0.489	0.562	0.533	0.565	0.568	0.446	0.562	0.536	0.557	0.455	0.488	0.552	0.487	0.515	0.5287
+ Lang Prior	bPQ	0.745	0.713	0.733	0.684	0.711	0.628	0.696	0.687	0.709	0.746	0.646	0.695	0.693	0.703	0.694	0.716	0.726	0.726	0.693	**0.7022**
	mPQ	0.577	0.539	0.628	0.558	0.536	0.516	0.577	0.548	0.564	0.588	0.467	0.579	0.549	0.582	0.473	0.501	0.584	0.514	0.542	**0.5484**

Table 8: **Automatic Instance Segmentation Performance across 14 Datasets.** mSA scores of UniNuc on the test splits of diverse histopathology datasets.

Metric	CPM15	CPM17	CoNSeP	CryoNuSeg	Lizard	LyNSec(H&E)	LyNSec(IHC)	MoNuSAC	MoNuSeg	NuClick	NuInsSeg	PanNuke	Puma	TNBC
mSA	0.5377	0.5157	0.2854	0.3227	0.4354	0.5080	0.4543	0.1948	0.3948	0.1450	0.3019	0.4821	0.5278	0.5245

The mSA is computed by averaging the accuracy score across multiple thresholds:

$$\text{mSA} = \frac{1}{|T|} \sum_{t \in T} \frac{TP(t)}{TP(t) + FP(t) + FN(t)} \tag{3}$$

where $T = \{0.5, 0.55, 0.6, \ldots, 0.95\}$. For each dataset, we report the mSA averaged over all images in the test set.

B.2. Automatic Instance Segmentation Metric

Consistent with CellViT (Hörst et al., 2024) and the PanNuke benchmark (Gamper et al., 2019b), we employ Panoptic Quality (PQ) to quantify instance segmentation performance. The PQ unifies detection and segmentation accuracy and is defined as:

$$PQ = \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{Detection Quality (DQ)}} \times \underbrace{\frac{\sum (y, \hat{y}) \in TP IoU(y, \hat{y})}{|TP|}}_{\text{Segmentation Quality (SQ)}} \tag{4}$$

where y denotes a ground-truth segment and $\hat{y}$ denotes a predicted segment. A pair $(y, \hat{y})$ is considered a unique match (True Positive) if $IoU(y, \hat{y}) > 0.5$. Based on this criterion, the set of segments is split into:

- True Positives (TP): Matched pairs of segments (correctly detected instances).
- False Positives (FP): Unmatched predicted segments (predicted instances with no ground truth).
- False Negatives (FN): Unmatched ground-truth segments (missed instances).

Intuitively, PQ decomposes into Detection Quality (DQ), which is analogous to the F_1 score, and Segmentation Quality (SQ), which measures the average IoU of matched segments. To

ensure a comprehensive evaluation, we report: Binary PQ (bPQ): Calculated by treating all nuclei as a single class (foreground vs. background). Multi-class PQ (mPQ): Calculated independently for each nuclei class and then averaged, providing a measure of class-specific performance.

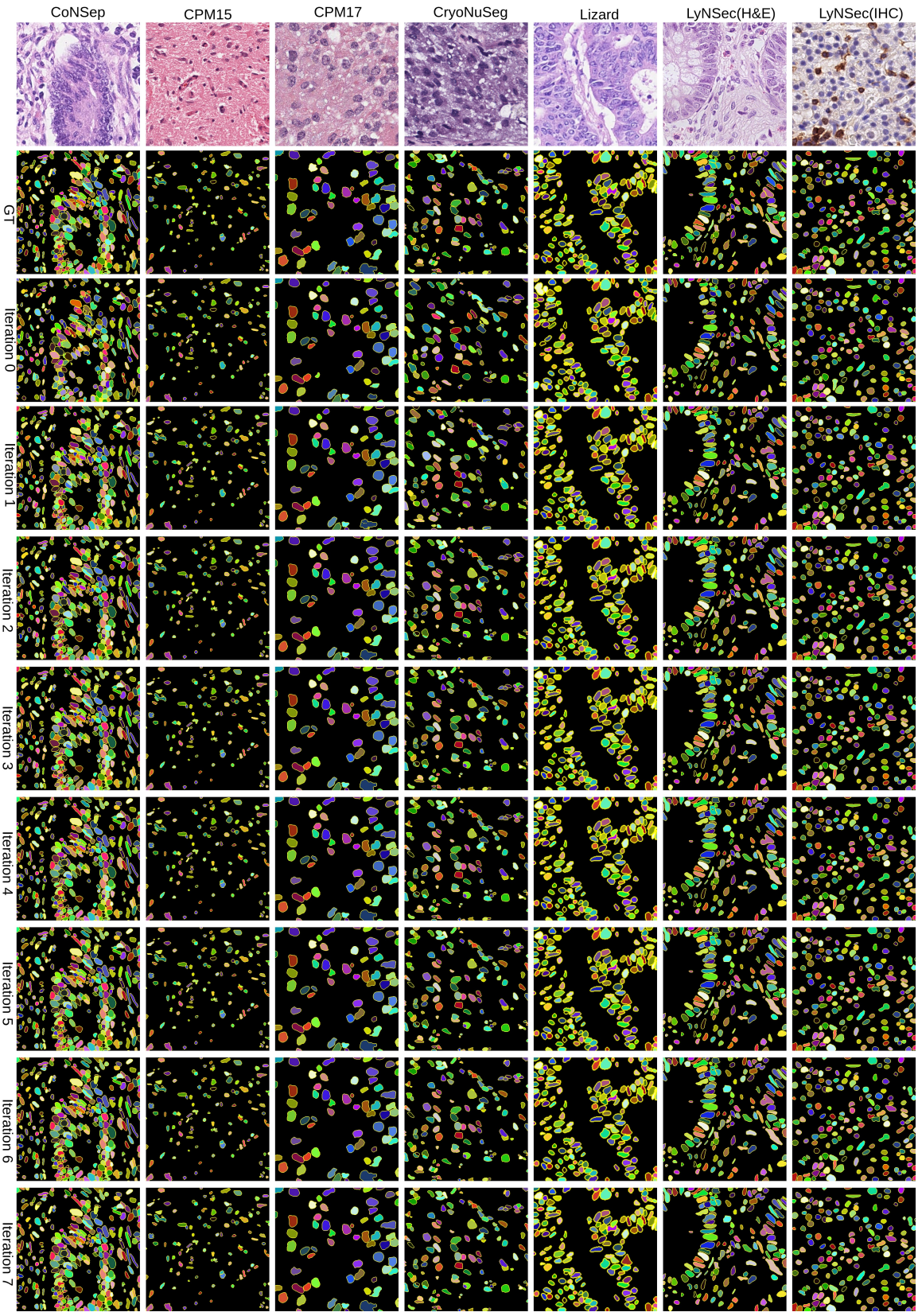

Figure 6: Qualitative plots for iterative interactive segmentation of UniNuc on CoNSep, CPM15, CPM17, CryoNuSeg, Lizard, LyNSec(H&E) and LyNSec(IHC). Top row: input/GT; subsequent rows: iterations 0–7. Best viewed zoomed in.

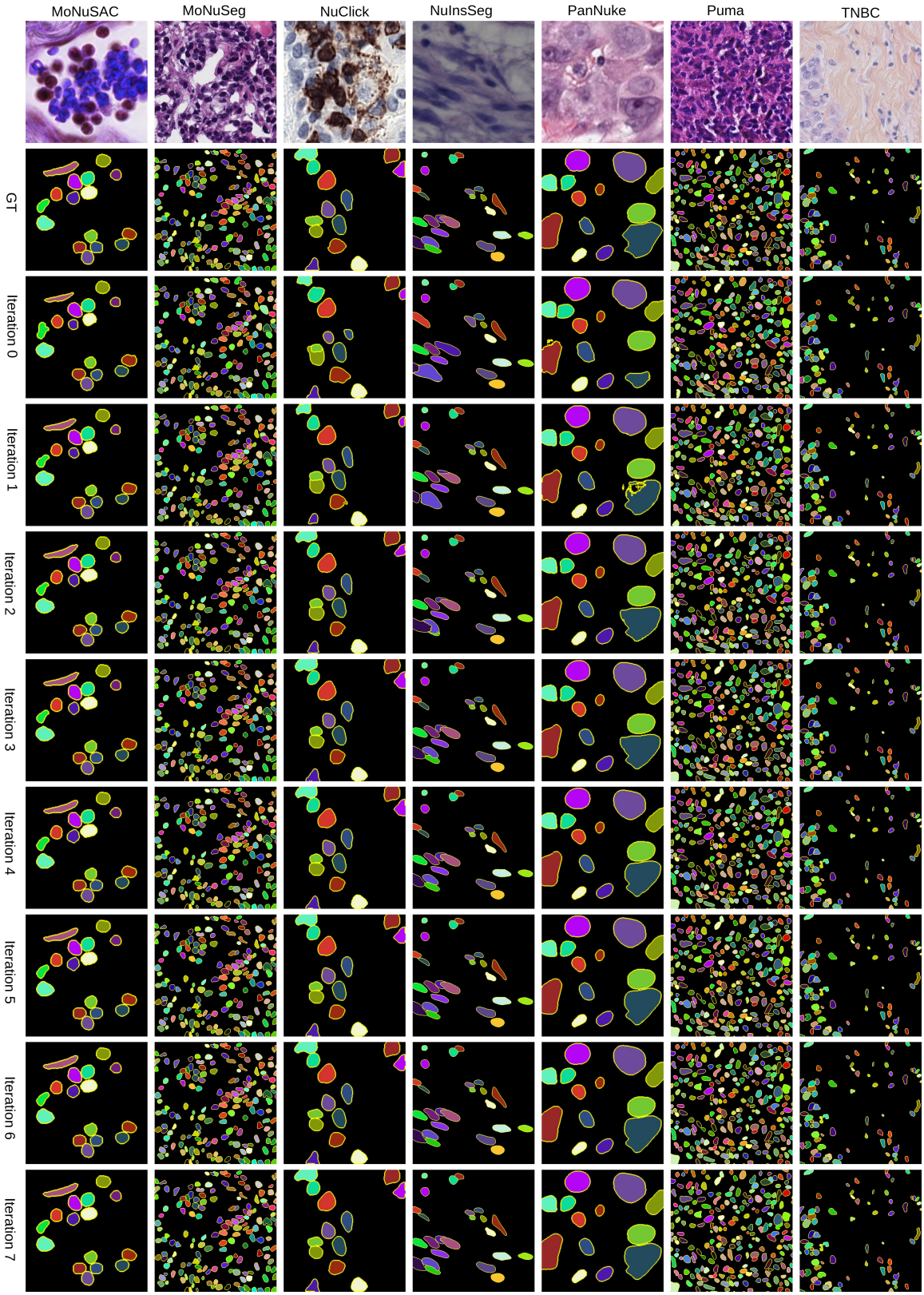

Figure 7: Qualitative plots for iterative interactive segmentation of UniNuc on MoNuSAC, MoNuSeg, NuClick, NuInsSeg, PanNuke, Puma and TNBC. Top row: input/GT; subsequent rows: iterations 0–7. Best viewed zoomed in.