S^2A-Net: A Semantic-Structural Aware Network for Nuclei Instance Segmentation and Classification

Can Jiang[1], Shaoguo Cui[1*], Guofen Wang[1]
[1]College of Computer and Information Science, Chongqing Normal University, Chongqing, China
916532442@qq.com, csg@cqnu.edu.cn, wanggf@cqnu.edu.cn

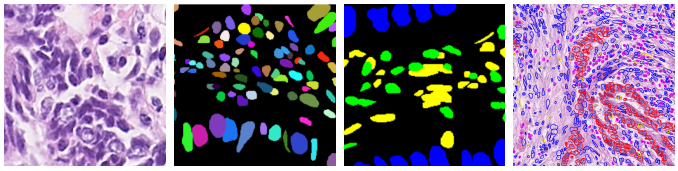
***Abstract*—Nuclear instance segmentation and classification are fundamental for quantifying the tumor microenvironment. However, strictly distinguishing adherent nuclei while identifying their fine-grained types remains challenging. Standard single-encoder architectures struggle to accommodate the conflicting feature requirements of fine-grained localization and high-level classification. Conversely, existing multi-branch designs tend to over-decouple these sub-tasks, breaking the intrinsic consistency between an instance's boundary and its semantic category. To address this, we propose S^2A-Net, featuring an asymmetric dual-stream architecture to explicitly disentangle task-specific features. We introduce the Dynamic Morpho-Semantic Rectifier(D-MSR) to inject attribute-level textual knowledge from medical Vision-Language Models into the visual pipeline. Acting as a semantic mediator, this module utilizes statistical priors to dynamically calibrate features for confusing cell subtypes without requiring explicit attribute annotations. Furthermore, a Directional Interaction Attention(DIA) module integrates direction-field constraints to explicitly disentangle adherent boundaries. Extensive experiments on the CoNSeP dataset and related benchmarks demonstrate that S^2A-Net achieves state-of-the-art performance and significantly enhances recognition accuracy for challenging cell categories.Code is available at https://github.com/jcan0809/S2Anet**

***Index Terms*—Digital pathology analysis, nuclei instance segmentation, nuclei classification, prompt learning.**

I. Introduction

Pathological image analysis of Hematoxylin and Eosin (H&E) stained Whole Slide Images serves as the gold standard for cancer diagnosis, grading, and prognosis estimation [1]. Within this domain, precise nuclear instance segmentation and classification are pivotal for quantifying the tumor microenvironment [2]. However, effectively parsing these microscopic structures in complex clinical scenarios remains a formidable challenge [3]. As illustrated in Fig. 1, colorectal adenocarcinoma tissues often exhibit highly dense nuclear distributions with severe overlapping boundaries that complicate instance separation, while the significant morphological heterogeneity among cell types compounds the difficulty of fine-grained classification.

Despite substantial progress driven by deep learning, achieving high-precision instance segmentation and fine-grained classification simultaneously is constrained by dual limitations in structural delineation and semantic recognition. Regarding structural modeling, standard methods ranging from general baselines like U-Net [4] and Mask R-CNN [5] to domain-specific architectures like HoVer-Net [6], Dist [1], and CD-Net [7] primarily focus on geometric delineation via pixel-wise masks or boundary refinement via distance regression, often treating classification as a secondary task. While recent multi-branch frameworks like SMILE [8] attempt to decouple these tasks, they typically rely on a symmetric, lightweight backbone for all branches. As discussed, this leads to either insufficient capacity for dense topological separation or computational redundancy for classification, while potentially breaking the spatial homology between boundary and category. Although recent advancements like ANet [9] and WeaveSeg [10] have introduced attention mechanisms and frequency-aware refinement respectively to sharpen feature details, these architectures are intrinsically limited by their reliance on visual information alone. Consequently, they struggle to capture high-level semantic priors that are essential for distinguishing morphologically similar subtypes.



(a) (b) (c) (d)

Fig. 1: **Challenges in Nuclei segmentation and classification**: (a) A representative colorectal adenocarcinoma image; (b) Tightly packed and crowded nuclear distribution; (c) Severely overlapping instances and indistinct contours; (d) Morphologically varied cell types and categories.

Regarding semantic guidance, precise recognition is bottlenecked by the scarcity of fine-grained annotations and rigid prior knowledge [11]. In recent years, Vision-Language Models such as CLIP [12] and PLIP [13] have emerged as powerful paradigms for learning transferable representations. While prompt-driven architectures like SAM [14], SAM2 [15] and CA-SAM2 [16] have revolutionized segmentation by incorporating contextual information, existing adaptations in computational pathology predominantly rely on static prompt learning [17]. Injecting fixed class names into the network fails to capture the drastic morphological variations of cellular instances within the same category. Consequently, how to flexibly integrate dynamic textual knowledge into visual feature learning without relying on explicit attribute annotations

* Shaoguo Cui is the corresponding author.

remains an open research question.

To address these limitations, we propose S^{2}A-Net, a Semantic-Structural Aware network featuring an asymmetric dual-stream architecture. Unlike conventional designs, we explicitly disentangle the feature extraction process to align with the specific complexity of each sub-task. The main contributions of this work are summarized as follows:

- We propose S^{2}A-Net, an asymmetric dual-stream architecture that strategically resolves the feature conflict between geometric delineation and semantic recognition by allocating backbone capacity according to task complexity.
- We introduce the Dynamic Morpho-Semantic Rectifier (D-MSR), which bridges the gap between visual features and linguistic knowledge by leveraging PLIP-derived priors to calibrate instance-level representations for confusing subtypes.
- We design the Directional Interaction Attention (DIA) module to explicitly incorporate topological gradient flows into the attention mechanism, thereby sharpening adherent boundaries in dense tissue regions.
- We demonstrate through extensive experiments on CoNSeP, MoNuSeg, and CPM17 that S^{2}A-Net establishes new state-of-the-art performance in both segmentation accuracy and fine-grained classification.

The data and code used in this study will be made publicly available upon acceptance of this paper.

II. Methods

In this section, we present the overall pipeline of S^{2}A-Net, visually demonstrated in Fig. 2, followed by an in-depth dissection of its key building blocks: the Dynamic Morpho-Semantic Rectifier (D-MSR) and the Directional Interaction Attention (DIA).

A. Asymmetric Dual-stream Architecture

To mitigate the inherent feature conflict between fine-grained geometric boundaries and high-level semantic categories, we construct an asymmetric dual-branch encoder that allocates computational resources based on task complexity. For the Structure Stream, we employ the high-capacity ResNet50 backbone. Nuclear instance segmentationan intrinsically dense prediction task characterized by severe cell adhesiondemands the stronger non-linear representational capabilities of a deeper architecture to resolve topological ambiguities and regress precise distance maps. In parallel, for the Semantic Stream, we utilize the lighter ResNet34. As empirically demonstrated by SMILE, a ResNet34-based encoder is sufficient to capture the discriminative features required for classification. This asymmetric design achieves an optimal balance, prioritizing geometric fidelity without incurring unnecessary computational overhead for semantic recognition.

B. Dynamic Morpho-Semantic Rectifier

To bridge the semantic gap between the text encoder and the visual streams, we propose the D-MSR module. As shown in Fig. 2(b), this module does not merely align features but dynamically rectifies visual representations using text-derived priors through a dual-branch interaction mechanism. Note that in Fig. 2(a), this central interaction unit is denoted as D-MSR.
Attribute Construction and Pseudo-label Generation. To supervise the semantic alignment in D-MSR without manual annotation, we construct a hierarchical Prompt Bank covering four dimensions: density, size, shape, and boundary complexity. During training, we generate instance-level pseudo-labels by quantifying morphometric descriptors from the ground truth masks. Specifically, Density and Size are determined by the instance count and mean pixel area within the patch, respectively. Shape categories are mapped using a combination of Aspect Ratio and Solidity , effectively distinguishing between spherical, spindle, and irregular morphologies. Similarly, Boundary smoothness is quantified using the Roughness metric, which is derived from the perimeterarea relation $P^2/4\pi A$. These statistical priors are aggregated into a multi-hot vector $\mathbf{y}_{attr}$ to optimize the attribute regularization loss $\mathcal{L}_{attr}$.
Dynamic Attribute Generation. Let $\mathbf{F}_{sem} \in \mathbb{R}^{C\times H\times W}$ denote the semantic features and $\mathbf{T}_{stat} \in \mathbb{R}^{K\times D}$ represent the static attribute embeddings encoded by the frozen PLIP text encoder from our Prompt Bank. To capture instance-specific variations, we formulate the dynamic embedding generation as:

$$\mathbf{T}_{dyn} = \mathbf{T}_{stat} + \alpha \cdot \mathcal{M}_{mlp}(\text{GAP}(\mathbf{F}_{sem})) \tag{1}$$

where $\mathcal{M}_{mlp}$ is a lightweight projection head, GAP$(\cdot)$ denotes global average pooling, and α is a learnable scaling factor initialized to 0.
Weakly-supervised Attribute Alignment. To ensure the generated $\mathbf{T}_{dyn}$ semantically matches the visual content, we compute an alignment score matrix $\mathbf{S}_{attr}$. As shown in the "R" block of Fig. 2(b), this is formulated as a relation operation:

$$\mathbf{S}_{attr} = \sigma\left(\frac{\mathbf{F}_{sem}^T \mathbf{W}_r \mathbf{T}_{dyn}}{\sqrt{d}}\right) \tag{2}$$

where $\mathbf{W}_r$ is a learnable projection matrix. This score is directly supervised by $\mathcal{L}_{attr}$ using the generated pseudo-labels $\mathbf{y}_{attr}$ to enforce semantic consistency before feature rectification.
Dual-Stream Rectification. The dynamic attributes are then injected into the visual streams. For the semantic branch, we employ a 1+Sigmoid gating mechanism to calibrate feature intensity:

$$\mathbf{F}'_{sem} = \mathbf{F}_{sem} \odot (\mathbf{1} + \sigma(\mathcal{F}_{1\times1}(\mathbf{T}_{dyn} \oplus \mathbf{F}_{sem}))) \tag{3}$$

For the structure branch, the SpatialGate projects priors into 2D space to refine boundaries:

$$\mathbf{F}'_{str} = (1-\beta)\mathbf{F}_{str} + \beta\,(\mathbf{F}_{str} \odot \mathcal{G}_{spa}(\mathbf{T}_{dyn})) \tag{4}$$

where $\oplus$ denotes broadcasting and concatenation, and β is a learnable fusion coefficient. Crucially, acting as a semantic

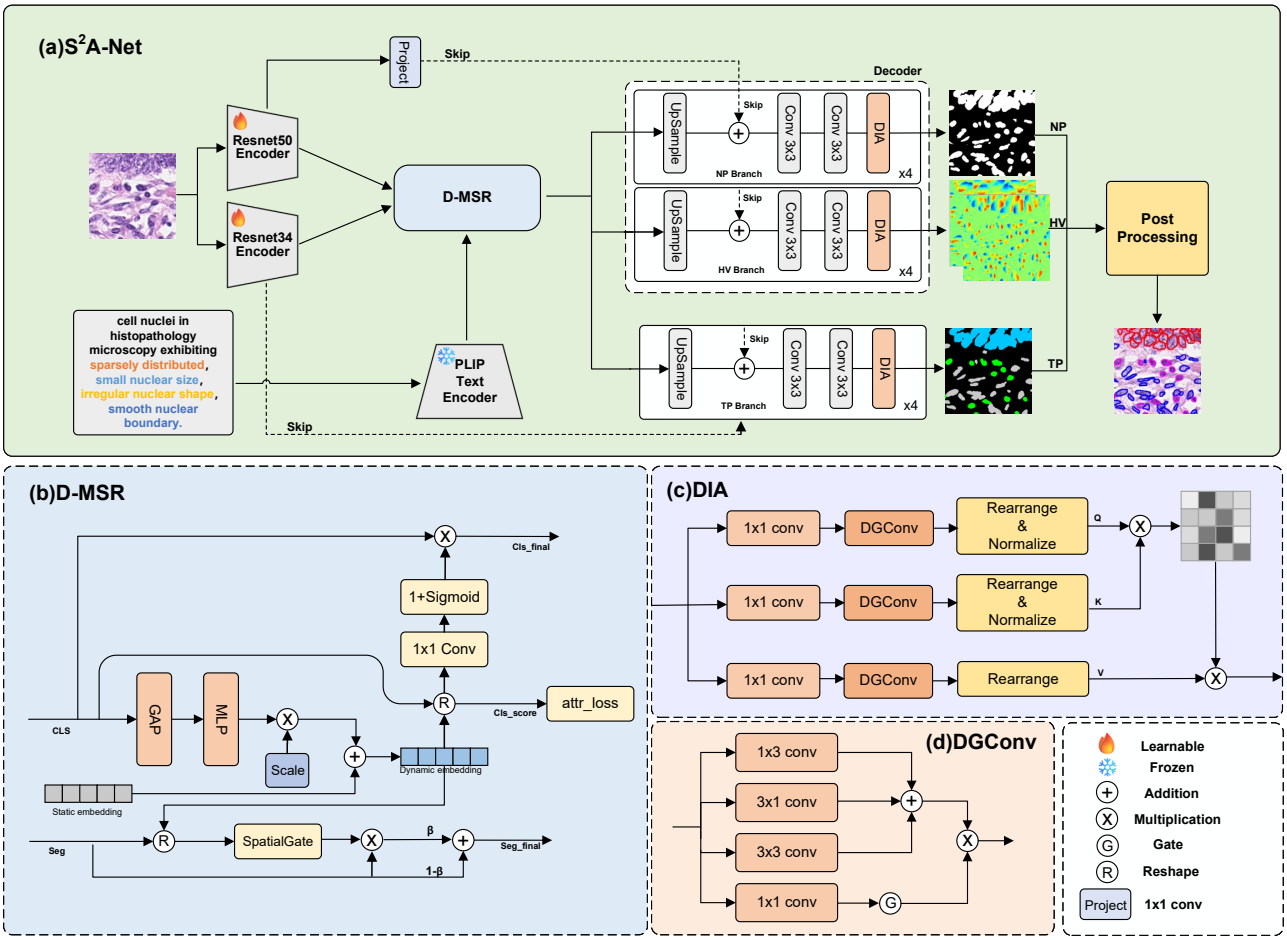


Fig. 2: Overall architecture of S^{2}A-Net and its key components. (a) The S^{2}A-Net framework features an asymmetric dual-stream encoder (ResNet50/34) and incorporates PLIP-derived textual priors to guide the decoding process with Directional Interaction Attention (DIA). (b) Dynamic Morpho-Semantic Rectifier (D-MSR) utilizing textual embeddings for instance-level feature calibration. (c) Internal structure of the DIA module designed to disentangle adherent boundaries via directional fields. (d) Deformable Group Convolution (DGConv) employing a multi-kernel strategy to capture irregular nuclear morphology.

mediator, D-MSR re-aligns the decoupled streams, restoring the intrinsic spatial homology between instance boundaries and their semantic categories.

C. Directional Interaction Attention

Standard convolutions possess isotropic receptive fields, which are ill-suited for separating adherent nuclei with complex boundaries. To explicitly model the topological gradient flow, we propose the DIA module. As illustrated in Fig. 2(c) and (d), DIA leverages a Deformable Group Convolution (DGConv) strategy to embed direction-aware geometric constraints into a self-attention mechanism.

Multi-Directional Perception via DGConv. The core of DIA is the DGConv unit. Unlike standard convolution, DGConv aggregates features from heterogeneous kernels to perceive anisotropic gradients. Let $\mathbf{X}$ be the input feature, the output of DGConv is defined as the summation of multi-branch transformations:

$$\mathrm{DG}(\mathbf{X}) = \sum_{k \in \mathcal{K}} (\mathbf{W}_k * (\mathbf{X} + \Delta\mathbf{p}_k)) \cdot \mathbf{m}_k \tag{5}$$

where $\mathcal{K} = \{1 \times 3, 3 \times 1, 3 \times 3\}$ represents the kernel set, and $\Delta\mathbf{p}_k, \mathbf{m}_k$ are the learned offsets and modulation masks for deformable convolution. This formulation explicitly models the vertical, horizontal, and omnidirectional topological flows.

Direction-Aware Attention Mechanism. We integrate DGConv into the QKV attention framework. The input features are projected into query, key, and value spaces via DGConv layers: $\mathbf{Q} = DG(\mathbf{X})\mathbf{W}_q$, $\mathbf{K} = DG(\mathbf{X})\mathbf{W}_k$, and $\mathbf{V} = DG(\mathbf{X})\mathbf{W}_v$.

To resolve the formatting issue of the long attention equation, we formulate the final output $\mathbf{F}_{out}$ as:

$$\mathbf{F}_{out} = Softmax\left(\frac{DG(\mathbf{Q}) \cdot DG(\mathbf{K})^T}{\sqrt{d_k}}\right) \cdot DG(\mathbf{V}) \tag{6}$$

By replacing standard linear projections with direction-aware $DG(\cdot)$ operations, the attention map $\mathbf{A}_{dir}$ explicitly encodes

topological relationships, effectively disentangling adherent instances.

D. Loss Functions

The S^{2}A-Net is optimized in an end-to-end manner via a multi-task objective function that jointly supervises foreground segmentation, instance regression, semantic classification, and attribute alignment. The total loss $\mathcal{L}_{total}$ is formulated as a weighted summation of four components:

$$\begin{aligned} \mathcal{L}_{total} &= \mathcal{L}_{NS} + \mathcal{L}_{NI} + \mathcal{L}_{NC} + \lambda_{attr}\mathcal{L}_{attr} \\ where \quad \mathcal{L}_{NS} &= \lambda_a\mathcal{L}_{ce} + \lambda_b\mathcal{L}_{dice} \\ \mathcal{L}_{NI} &= \lambda_c\mathcal{L}_{mse} + \lambda_d\mathcal{L}_{msge} \\ \mathcal{L}_{NC} &= \lambda_e\mathcal{L}_{cs} + \lambda_f\mathcal{L}_{ad} \end{aligned} \tag{7}$$

In this formulation, $\lambda_{a...f}$ and λ_{attr} serve to balance the multi-task objective. Following HoVer-Net [6], we employ a hybrid of Cross-Entropy and Dice loss ($\mathcal{L}_{NS}$) for segmentation, alongside Mean Squared Error and Gradient Error ($\mathcal{L}_{NI}$) for distance map regression. For classification ($\mathcal{L}_{NC}$), we adopt the cost-sensitive strategy from SMILE [8] to mitigate class imbalance. Crucially, to supervise the D-MSR module, we introduce an Attribute Regularization term ($\mathcal{L}_{attr}$). This term utilizes binary cross-entropy as a weak supervision signal to enforce explicit alignment between visual features and the linguistic priors derived from the text encoder.

III. Experiments

A. Settings

TABLE I: Summary of the Datasets Used in Our Experiments.

Dataset	Task	Nuclei	Mag.	Organ
CoNSeP [6]	S.+C.	24,319	40×	Colon
MoNuSeg [18]	Seg.	21,623	40×	Multi.
CPM-17 [19]	Seg.	7,570	20/40×	Multi.

a)Datasets

We evaluate our method on three public benchmarks: CoNSeP [6], MoNuSeg [18], and CPM17 [19], as summarized in Table I. CoNSeP serves as the primary dataset for joint segmentation and classification, comprising 41 H&E-stained colorectal adenocarcinoma tiles annotated with four fine-grained nuclear categories including Epithelial, Inflammatory, Spindle, and Miscellaneous. To assess segmentation generalization across diverse tissue types, we utilize the multi-organ MoNuSeg and CPM17 datasets derived from TCGA, which contain 30 and 32 images respectively. We strictly adhere to the official train-test splits for all datasets to ensure fair comparison.

b) Evaluation Metrics

To comprehensively assess performance, we employ standard quantitative metrics established in computational pathology. Segmentation accuracy is measured using the Dice Coefficient to quantify pixel-level overlap and the Aggregated Jaccard Index, abbreviated as AJI, for object-level assessment. The latter is particularly critical as it rigorously penalizes under-segmentation. Panoptic performance is evaluated via Panoptic Quality, denoted as PQ. Defined as the product of Segmentation Quality and Detection Quality, this metric offers a holistic assessment of boundary delineation and detection performance. For the fine-grained CoNSeP dataset, we additionally report Detection F1, denoted as F_d, and class-specific Classification F1, denoted as F_c, to validate multi-class recognition capabilities across all annotated cell types.

c) Implementation Details

S^{2}A-Net was implemented using PyTorch on a single NVIDIA RTX A6000 GPU. The ResNet34 backbone was initialized with histology-specific Cerberus weights while the text encoder remained frozen. We employed the AdamW optimizer with an initial learning rate of 1×10^{-4} managed by a cosine annealing schedule. A two-stage training strategy was adopted consisting of a 50-epoch warm-up phase where the backbone remains frozen with a batch size of 8, followed by a 70-epoch fine-tuning phase utilizing an unfrozen backbone and a batch size of 4. Loss weights were empirically set to 0.5 for attribute regularization and 1.0 for all other tasks. The morphometric thresholds for pseudo-label generation were empirically determined based on the statistical distribution of the CoNSeP training set. Regarding data preparation, we utilized a sliding window strategy to extract standardized 256×256 patches for detail preservation, further enhancing robustness through rigorous augmentations such as random flipping, rotation, and Gaussian noise.

B. Results and comparison

1) Nuclei Instance Segmentation and Classification Tasks:

As summarized in Table II, S^{2}A-Net achieves comprehensive state-of-the-art performance on the CoNSeP dataset. In segmentation, our method notably outperforms the recent CP-Mamba by 3.6% in AJI and 3.9% in PQ, underscoring the efficacy of our structure-aware design in delineating dense boundaries. Regarding classification, while SMILE retains a marginal edge in the epithelial class, S^{2}A-Net demonstrates superior generalization across all other categories. Crucially, in the challenging miscellaneous category (F_c^m), we outperform CP-Mamba and SMILE by significant margins (48.8% vs. 41.0% and 37.8%), confirming that incorporating textual priors effectively mitigates semantic ambiguity in morphologically diverse subtypes. Qualitative visualizations are provided in the bottom row of Fig. 3.

2) Nuclei Semantic and Instance Segmentation Tasks:

Tables III and IV present the evaluation on CPM-17 and MoNuSeg datasets. On CPM-17, our method achieves leading performance with the highest Dice (88.4%), SQ (82.4%), and PQ (72.5%). Although our AJI (73.8%) is slightly lower than CA-SAM2, it significantly outperforms established baselines like HoVer-Net. Extending to MoNuSeg, S^{2}A-Net demonstrates strong generalization, securing the best AJI (66.9%), DQ (83.9%), and PQ (65.0%). Notably, our AJI surpasses WeaveSeg (64.2%) and HoVer-Net (66.1%), highlighting the distinct advantage of our asymmetric design in separating touching nuclei within dense tissues. Despite a marginally

TABLE II
Nuclear Segmentation and Classification Results on the CoNSeP Dataset. F_d denotes F_1 score for nuclear detection, F_c^* is classification score for each nucleus type.

Method	Segmentation Metric					Classification Metric					Publications
	Dice	AJI	DQ	SQ	PQ	F_d	F_c^e	F_c^i	F_c^s	F_c^m	
U-Net [4]	0.724	0.482	0.488	0.671	0.328	0.629	0.410	0.550	0.481	0.000	MICCAI'2015
Mask-RCNN [5]	0.740	0.474	0.619	0.740	0.460	0.692	0.595	0.590	0.520	0.098	ICCV'2017
DIST [1]	0.804	0.502	0.544	0.728	0.398	0.714	0.616	0.609	0.527	0.172	TMI'2018
HoVer-Net [6]	0.853	0.571	0.702	0.778	0.547	0.748	0.635	0.631	0.566	0.426	MIA'2019
Full-Net [20]	0.847	0.479	0.642	0.776	0.498	0.720	0.580	0.560	0.530	0.150	MICCAI'2019
CDNet [7]	0.835	0.541	0.652	0.771	0.505	0.710	0.574	0.519	0.501	0.300	ICCV'2021
SMILE [8]	0.833	0.544	0.678	0.766	0.521	0.763	**0.668**	0.624	0.578	0.378	MIA'2023
ANet [9]	0.850	0.563	0.703	**0.779**	0.549	0.743	0.625	0.613	0.579	0.351	MDPI'2024
CP-Mamba [21]	0.819	0.539	0.658	0.776	0.514	0.753	0.657	0.632	0.571	0.410	AAAI'2025
WeaveSeg [10]	0.852	0.563	-	-	0.548	0.743	-	-	-	-	ICCV'2025
Ours	**0.855**	**0.575**	**0.711**	0.776	**0.553**	**0.772**	0.658	**0.645**	**0.603**	**0.488**	-

TABLE III
Nuclear Segmentation Results on the CPM-17 Dataset.

Method	Dice	AJI	DQ	SQ	PQ	Publications
U-Net [4]	0.813	0.643	0.778	0.734	0.578	MICCAI'2015
Mask-RCNN [5]	0.850	0.684	0.848	0.792	0.674	ICCV'2017
DIST [1]	0.826	0.616	0.663	0.754	0.504	TMI'2018
HoVer-Net [6]	0.869	0.705	0.854	0.814	0.697	MIA'2019
Full-Net [20]	0.878	0.670	0.849	0.806	0.684	MICCAI'2019
CDNet [7]	0.880	0.733	0.873	0.812	0.710	ICCV'2021
SMILE [8]	0.877	0.716	0.858	0.810	0.696	MIA'2023
ANet [9]	0.881	0.724	0.866	0.815	0.706	MDPI'2024
CA-SAM2 [16]	0.881	**0.746**	**0.881**	0.819	0.723	MICCAI'2025
Ours	**0.884**	0.738	0.880	**0.824**	**0.725**	-

TABLE IV
Nuclear Segmentation Results on the MoNuSeg Dataset.

Method	Dice	AJI	DQ	SQ	PQ	Publications
U-Net [4]	0.758	0.556	0.691	0.690	0.478	MICCAI'2015
Mask-RCNN [5]	0.760	0.546	0.522	0.700	0.809	ICCV'2017
DIST [1]	0.789	0.559	0.601	0.732	0.443	TMI'2018
HoVer-Net [6]	0.821	0.661	0.771	0.763	0.590	MIA'2019
Full-Net [20]	0.820	0.600	0.771	**0.790**	0.615	MICCAI'2019
CDNet [7]	0.826	0.630	0.802	**0.790**	0.634	ICCV'2021
SMILE [8]	0.782	0.576	0.709	0.749	0.533	MIA'2023
ANet [9]	0.825	0.618	0.766	0.765	0.587	MDPI'2024
WeaveSeg [10]	**0.836**	0.642	-	-	0.637	ICCV'2025
CA-SAM2 [16]	0.810	0.619	0.760	0.759	0.579	MICCAI'2025
Ours	0.830	**0.669**	**0.839**	0.775	**0.650**	-

lower Dice compared to WeaveSeg, the overall dominance in instance-level metrics confirms the reliability of our approach for precise quantitative analysis. Visualizations are provided in the top rows off Fig. 3.

C. Ablation Studies

Table V validates our architectural choices. The asymmetric encoder utilizing a ResNet50 structure branch elevates the AJI from 53.2% in the symmetric ResNet34 baseline to 55.6%, confirming that the structural stream demands higher capacity and stronger non-linear representations to resolve dense topology. Regarding key modules, the D-MSR targets semantic ambiguity among confusing subtypes by increasing the F_c^m score from 35.5% to 43.9% via injected morphometric priors, while the DIA improves the AJI to 56.6% by modeling anisotropic gradient flows to explicitly disentangle

TABLE V
Ablation Study of Key Components on the CoNSeP Dataset.

(a) Segmentation Metrics

Method	Dice	AJI	DQ	SQ	PQ
Single Img. Enc.	0.838	0.530	0.657	0.766	0.505
SMILE Img. Enc.	0.836	0.532	0.658	0.764	0.503
Dual-stream Img. Enc.	0.841	0.556	0.688	0.773	0.532
DS Img. Enc. + D-MSR	0.847	0.562	0.701	0.774	0.543
DS Img. Enc. + DIA	0.845	0.566	0.703	0.773	0.544
DS Img. Enc. + Both	**0.855**	**0.575**	**0.711**	**0.776**	**0.553**

(b) Classification Metrics

Method	F_d	F_c^e	F_c^i	F_c^s	F_c^m
Single Img. Enc.	0.743	0.622	0.628	0.566	0.281
SMILE Img. Enc.	0.748	0.648	0.603	0.562	0.270
Dual-stream Img. Enc.	0.755	0.628	0.604	0.564	0.355
DS Img. Enc. + D-MSR	0.768	0.636	0.612	0.575	0.439
DS Img. Enc. + DIA	0.765	0.613	0.589	0.574	0.425
DS Img. Enc. + Both	**0.772**	**0.658**	**0.645**	**0.603**	**0.488**

Note: **DS Img. Enc.** denotes Dual-stream Image Encoder. *Single Img. Enc.* utilizes ResNet-50, while *SMILE Img. Enc.* employs ResNet-34. Our full model (+ *Both*) integrates D-MSR and DIA modules.

adherent boundaries. Integrating both components yields peak performance with an AJI of 57.5% and an F_c^m of 48.8%, illustrating the complementary nature of semantic calibration and boundary refinement in resolving feature conflict.

IV. Conclusion

In this work, we proposed S^{2}A-Net to address the conflicting feature requirements of fine-grained nuclear segmentation and high-level classification. By employing an asymmetric dual-stream architecture that aligns model capacity with task complexity, we explicitly disentangled structural delineation from semantic recognition. Furthermore, the integration of Vision-Language Model priors via the Dynamic Morpho-Semantic Rectifier and topological constraints via Directional Interaction Attention significantly enhances robustness in complex tissue regions. Experimental results confirm that S^{2}A-Net establishes new state-of-the-art performance. Future work will focus on validating this asymmetric framework on broader

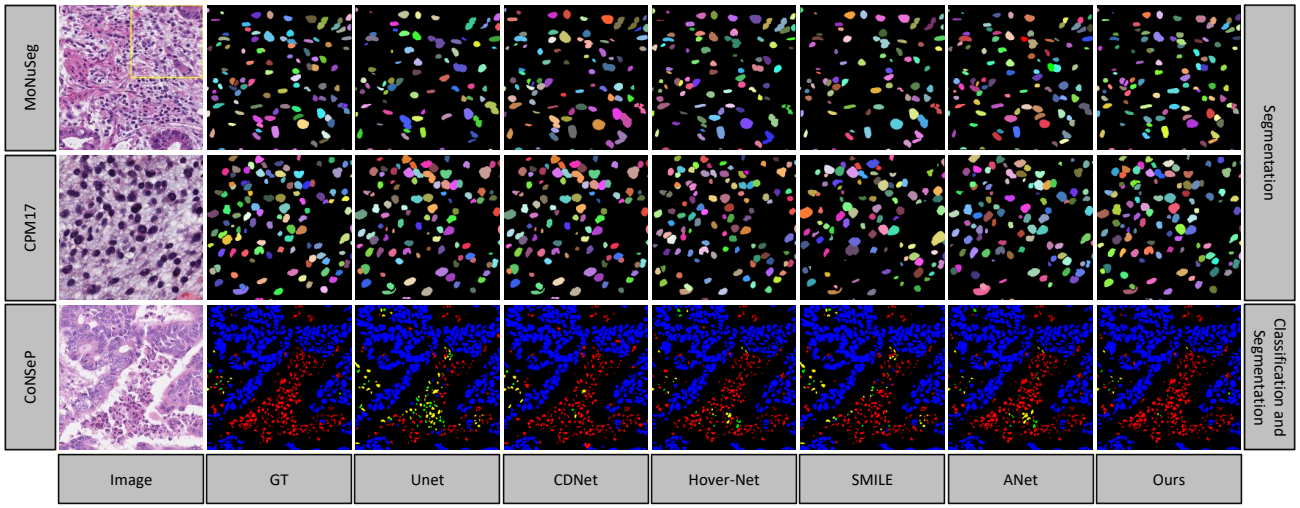


Fig. 3: Visualization results on MoNuSeg, CPM-17 and CoNSeP datasets. For instance segmentation (top 2 rows), different colors of nuclear boundaries represent distinct instances, while for classification results (bottom row), varying colors indicate different types. Regions of evident improvements are enlarged via yellow boxes to show better details.

histopathology tasks and exploring more effective mechanisms to integrate multi-modal knowledge.

V. Acknowledgments

Key Project of Chongqing Natural Science Foundation Innovation Development Joint Fund (CSTB2025NSCQ-LZX0073)

References

[1] P. Naylor, M. Laé, F. Reyal, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 448–459, 2018.

[2] N. Zamanitajeddin, M. Jahanifar, M. Bilal, M. Eastwood, and N. Rajpoot, "Social network analysis of cell networks improves deep learning for prediction of molecular pathways and key mutations in colorectal cancer," *Medical Image Analysis*, vol. 93, p. 103071, 2024.

[3] D. Liu, D. Zhang, Y. Song, H. Huang, and W. Cai, "Panoptic feature fusion net: a novel instance segmentation paradigm for biomedical and biological images," *IEEE Transactions on Image Processing*, vol. 30, pp. 2045–2059, 2021.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.

[6] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical image analysis*, vol. 58, p. 101563, 2019.

[7] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "Cdnet: Complementary depth network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3376–3390, 2021.

[8] X. Pan, J. Cheng, F. Hou, R. Lan, C. Lu, L. Li, Z. Feng, H. Wang, C. Liang, Z. Liu *et al.*, "Smile: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations," *Medical Image Analysis*, vol. 88, p. 102867, 2023.

[9] M. Kadaskar and N. Patil, "Anet: Nuclei instance segmentation and classification with attention-based network," *SN Computer Science*, vol. 5, no. 4, p. 348, 2024.

[10] J. Li, H. Wu, and J. Qin, "Weaveseg: Iterative contrast-weaving and spectral feature-refining for nuclei instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 21 984–21 993.

[11] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 113–19 122.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[13] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[15] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[16] H. Huang, H. He, L. Xu, X. Zhu, S. Feng, and G. Fu, "Ca-sam2: Sam2-based context-aware network with auto-prompting for nuclei instance segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 86–95.

[17] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[18] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu *et al.*, "A multi-organ nucleus segmentation challenge," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1380–1391, 2019.

[19] Q. D. Vu, S. Graham, T. Kurc, M. N. N. To, M. Shaban, T. Qaiser, N. A. Koohbanani, S. A. Khurram, J. Kalpathy-Cramer, T. Zhao *et al.*, "Methods for segmentation and classification of digital microscopy tissue images," *Frontiers in bioengineering and biotechnology*, vol. 7, p. 53, 2019.

[20] H. Qu, Z. Yan, G. M. Riedlinger, S. De, and D. N. Metaxas, "Improving nuclei/gland instance segmentation in histopathology images by full resolution neural network and spatial constrained loss," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 378–386.

[21] Y. Zhang, Z. Fang, Y. Wang, L. Zhang, X. Guan, and Y. Zhang, "Category prompt mamba network for nuclei segmentation and classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 284–10 292.