S^{3}H-Net: Injecting Hybrid Semantic-Spatial Guidance via Superpixel Hypergraphs for Medical Image Segmentation

Zongjian Yang[1], Lanxiang Ma[1], Zeli Gao[2], Yuqi Zhang[1], and Jiquan Ma[1,3](✉)

[1] School of Computer and Big Data, Heilongjiang University, Harbin, China
majiquan@hlju.edu.cn
[2] School of Mathematics and Statistics, Heilongjiang University, Harbin, China
[3] Hunan Provincial Key Laboratory of Multi-omics and Artificial Intelligence of Cardiovascular Diseases, University of South China, Hengyang, China

Abstract. Accurate segmentation of medical images remains a intricate challenge due to the complex geometry of target structures and their blurred boundaries. While recent approaches leverage topological and semantic information by introducing graph or hypergraph, they face two critical limitations: i) fixed patch nodes inherently ignore object boundaries, causing semantic ambiguity by mixing heterogeneous features; and ii) restricting reasoning to bottlenecks or skip connections fails to fully inject global contexts for fine-grained segmentation. To address these issues, we propose **S^{3}H-Net**, a novel architecture that integrates **S**uperpixel-based **S**emantic-**S**patial **H**ypergraphs. Specifically, distinct from patch-based methods, we utilize a boundary-aware Simple Linear Iterative Clustering (SLIC) algorithm for node construction, encouraging contour-aware regional grouping and alleviating heterogeneous feature mixing. Furthermore, we design Hybrid Semantic-Spatial Hypergraph (HSSH) blocks to construct a parallel Semantic Encoder, which captures multi-scale high-order semantic guidance continuously. To explicitly bridge the semantic gap between the non-local hypergraph topology and local convolutional features, we introduce a Cross-Covariance Semantic Injection (CCSI) module. By exploiting inter-channel attention, this module effectively injects the reasoned global context into local features for comprehensive semantic awareness. Extensive experiments on five medical imaging modalities demonstrate that S^{3}H-Net significantly outperforms state-of-the-art methods. Our code is available on GitHub.

Keywords: Medical Image Segmentation · Semantic-Spatial Hypergraph · Superpixel · Semantic Injection

1 Introduction

Accurate segmentation of anatomical structures and lesions is a fundamental prerequisite for computer-aided diagnosis, treatment planning, and clinical monitoring. However, this task remains a formidable challenge due to the high variability

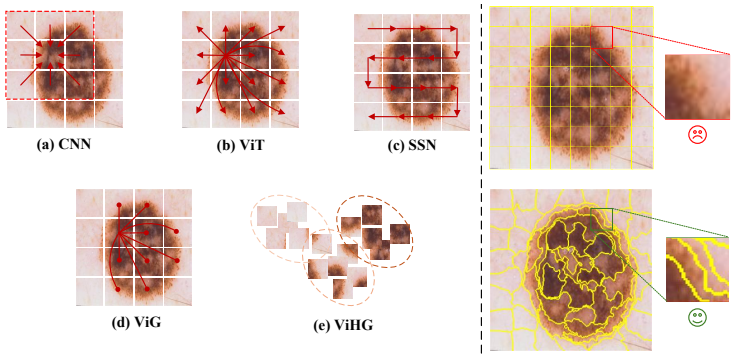

Fig. 1. Left: Comparison of visual modeling paradigms. (a-c) Existing methods distinct by local windows (CNN), global dense attention (ViT), or 1D scanning sequences (SSN). (d) ViG captures pairwise relations via graph edges. (e) ViHG constructs hyperedges to group multiple nodes sharing high-order semantics, offering a superior representation for structural understanding. **Right: Comparison of patch and superpixel.** Patch partition (top) ignores the image content and forcibly cuts off the object boundaries resulting in feature contamination. Superpixel (bottom) adapts to the anatomical boundaries adaptively, ensuring semantic consistency, meaning that each node is pure, providing precise structural guidance for segmentation.

in complex target geometries and the inherent ambiguity of blurred boundaries caused by scanning device and image reconstruction [23].

To capture contextual information for medical image segmentation, various modeling paradigms have been explored. As illustrated in Fig. 1(a-c), dominant architectures like CNNs [11], Vision Transformers (ViTs) [5], and State Space Models (SSMs) [19] primarily focus on grid-like or sequential feature aggregation. While effective for local texture or long-range dependencies, these methods often struggle to explicitly model the non-local structural topology and disjoint semantic connections inherent in medical images. Consequently, Vision Graphs (ViGs) [8,17] and Vision Hypergraphs (ViHGs) [4] have emerged as superior alternatives, offering a flexible mechanism to capture pairwise or high-order semantic correlations among irregular regions (Fig. 1d-e).

However, the potential of these graph-based models is severely limited by their node representation. Most approaches rigidly partition images into fixed square patches (Fig. 1 Right), which inherently disregards anatomical boundaries. This misalignment forces foreground and background features to mix within a single node, leading to semantic ambiguity that contaminates the subsequent topological reasoning. Compounding this issue is the suboptimal placement of the reasoning module. Predominant approaches typically restrict graph or hypergraph learning to the network's bottleneck [13] or simple skip connections [4,17]. While this strategy reduces computational overhead by operating at the low resolution, it fails to interact with multi-scale features. This late-stage interaction creates a significant semantic gap, where the abstract topological

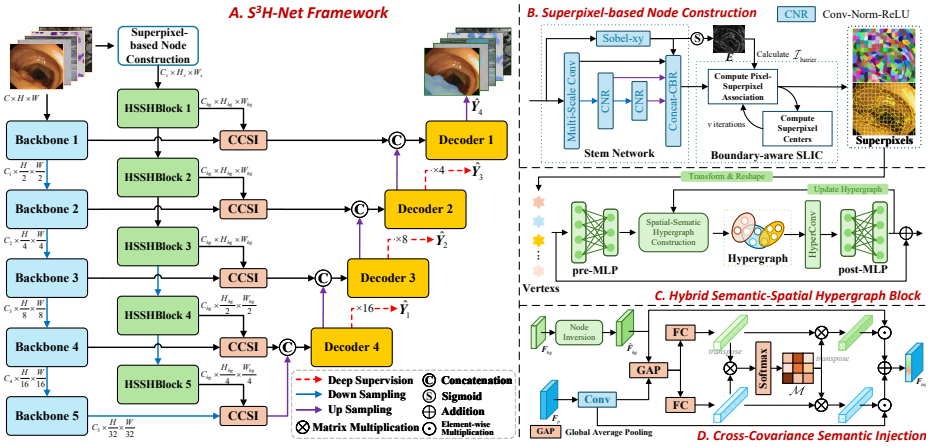

Fig. 2. The overview of proposed S^{3}H-Net.

information is too coarse to effectively guide the recovery of fine-grained boundaries in the shallow layers.

To tackle the aforementioned challenges, we propose **S^{3}H-Net**, a novel architecture that integrates **S**uperpixel-based **S**emantic-**S**patial **H**ypergraphs into a U-shaped **Net**work. First, distinct from patch-based approaches, we reformulate the classical SLIC algorithm [1] with boundary awareness for node construction. Specifically, we reformulate the SLIC distance metric by incorporating a barrier-based boundary term. This modification prioritizes structural continuity over mere color similarity, ensuring that the generated superpixels strictly adhere to anatomical contours (Fig. 1, Right bottom) and effectively eliminate semantic ambiguity at the source. Second, to capture high-order semantic correlations, we design Hybrid Semantic-Spatial Hypergraph (HSSH) blocks within a parallel Semantic Encoder. By constructing hybrid hyperedges based on spatial proximity and feature similarity, these blocks extract continuous semantic correlations across multiple scales. Finally, the Cross-Covariance Semantic Injection (CCSI) module bridges the topological-spatial gap. It projects reasoned hypergraph features back to the pixel space via node inversion, utilizing a channel-wise cross-covariance mechanism to explicitly inject global semantic guidance into local spatial representations. Extensive experiments on five modalities demonstrate that our S^{3}H-Net significantly outperforms state-of-the-art methods.

2 Methods

As shown in Fig. 2(A), S^{3}H-Net employs a dual-branch design, which parallels a standard CNN backbone with a Semantic Encoder based on HSSH blocks. The Semantic Encoder captures high-order semantic correlations, which are then hierarchically injected through the CCSI module to dynamically rectify local pixel-wise features.

2.1 Superpixel-based Node Construction

Differentiable Boundary-aware SLIC Algorithm. Unlike grid patches that disrupt anatomical structures, we employ a differentiable SLIC module following the multi-scale Stem Network to generate perceptually meaningful superpixels $\mathcal{S} \in \mathbb{R}^{C_s \times H_s \times W_s}$. To capture structural boundaries explicitly, the network utilizes a Sobel operator to build a probabilistic edge map $\boldsymbol{E} \in [0,1]^{H \times W}$ (Fig. 2(B)).

We adopt a differentiable soft association strategy [12] to generate the assignment matrix $\boldsymbol{Q} \in \mathbb{R}^{N \times M}$ linking N pixels to M superpixels. Each entry $q_{ps} = \frac{e^{-D_{ps}^2/\tau}}{\sum_k e^{-D_{pk}^2/\tau}}$ represents the association probability modulated by temperature $\tau = 0.01$. For a pixel p and superpixel center s, let $\boldsymbol{f}_p, \boldsymbol{f}_s$ and $\boldsymbol{x}_p, \boldsymbol{x}_s$ denote their feature vectors and spatial coordinates, respectively. The association cost D_{ps} is reformulated to incorporate spatial and boundary-aware constraints:

$$D_{ps} = \|\boldsymbol{f}_p - \boldsymbol{f}_s\|^2 + \underbrace{\lambda_{xy}\|\boldsymbol{x}_p - \boldsymbol{x}_s\|^2}_{\mathcal{R}_{\text{spatial}}} + \underbrace{\lambda_b \int_0^1 \boldsymbol{E}(\boldsymbol{x}_p + t(\boldsymbol{x}_s - \boldsymbol{x}_p))\, dt}_{\mathcal{I}_{\text{barrier}}}, \quad (1)$$

where λ_{xy} and λ_b balance two regularization terms. Specifically, $\mathcal{R}_{\text{spatial}}$ enforces compactness, while the barrier $\mathcal{I}_{\text{barrier}}$ integrates edge probability $\boldsymbol{E}$ along the trajectory between $\boldsymbol{x}_p$ and $\boldsymbol{x}_s$. To enhance precision, we approximate this integral via the composite trapezoidal rule:

$$\mathcal{I}_{\text{barrier}} \approx \frac{1}{2K}\left[\boldsymbol{E}(\boldsymbol{x}_0) + \boldsymbol{E}(\boldsymbol{x}_K) + 2\sum_{k=1}^{K-1} \boldsymbol{E}(\boldsymbol{x}_k)\right], \quad (2)$$

where $\boldsymbol{x}_k$ are $K = 8$ sampled points along the path. This mechanism ensures boundary adherence by penalizing superpixel leakage across semantic contours.

Node Inversion. To project the enhanced node features $\boldsymbol{F}_{hg} \in \mathbb{R}^{(H_{hg} \times W_{hg}) \times C_{hg}}$ back to the spatial scale of pixel features $\boldsymbol{F}_p \in \mathbb{R}^{(H_p \times W_p) \times C_p}$, we compute a new assignment matrix $\boldsymbol{Q}'$ derived from $\boldsymbol{F}_p$. We then perform *superpixel upsampling* [20] via matrix multiplication $\hat{\boldsymbol{F}}_{hg} = \boldsymbol{Q}'\boldsymbol{F}_{hg}$ to obtain the pixel-level map $\hat{\boldsymbol{F}}_{hg} \in \mathbb{R}^{(H_p \times W_p) \times C_{hg}}$. Unlike standard bilinear upsampling, this content-aware operation enforces intra-region semantic consistency, uniformly propagating global context within anatomical units to prevent cross-tissue ambiguity.

2.2 Hybrid Semantic-Spatial Hypergraph (HSSH)

To explicitly model complex anatomical topology and provide robust semantic guidance, we construct a hypergraph to capture high-order semantics within morphological superpixel nodes. Formally, a hypergraph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and is structurally represented by an incidence matrix $\boldsymbol{H} \in [0,1]^{|\mathcal{V}| \times |\mathcal{E}|}$, where each entry $\boldsymbol{H}_{ve}$ quantifies the affinity between the vertex $\boldsymbol{v}$ and the hyperedge $\boldsymbol{e}$ [7]. In our S^{3}H-Net, we initialize the vertex set $\mathcal{V}$ using SLIC-derived superpixels $\mathcal{S}$ as the basic units for subsequent semantic-spatial interaction.

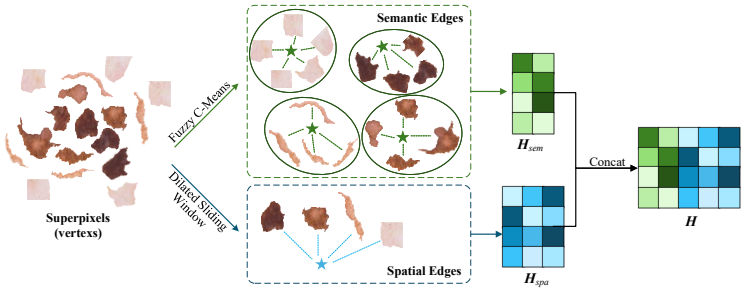

Fig. 3. Hybrid semantic-spatial hypergraph (HSSH) construction.

Hypergraph Construction. After initializing the vertex set with superpixels, we construct hyperedges from two complementary perspectives (Fig. 3): *semantic correlation* and *spatial connectivity*. For semantic hyperedges, we aim to capture long-range dependencies among disjoint regions. Instead of the standard k-Nearest Neighbors (kNN) approach, which is computationally expensive ($O(N^2)$) and sensitive to noise, we employ the Fuzzy C-Means (FCM) algorithm [9]. FCM projects superpixels into a probabilistic space, yielding a semantic incidence matrix $\boldsymbol{H}_{sem} \in [0,1]^{N\times K_{sem}}$ that represents the soft membership of each vertex to latent semantic centers. Complementarily, relying solely on semantic similarity may neglect local anatomical continuity. To address this, we construct spatial hyperedges using a dilated sliding window strategy under a stride of 2, generating a binary matrix $\boldsymbol{H}_{spa} \in \{0,1\}^{N\times K_{spa}}$ to explicitly encode local neighborhood structures. We deliberately avoid the dense local sliding window strategy [4] to prevent structural redundancy and reduce computational complexity. Finally, we fuse these views into a hybrid incidence matrix via concatenation, donated as $\boldsymbol{H} = [\boldsymbol{H}_{sem} \parallel \boldsymbol{H}_{spa}] \in \mathbb{R}_{\geq 0}^{N\times(K_{sem}+K_{spa})}$.

Hypergraph Convolution and Block Architecture. Based on the constructed $\boldsymbol{H}$, we employ the spectral hypergraph convolution [7]. Let $\boldsymbol{X}$ denote the input features. The propagation is formulated as:

$$\boldsymbol{Z} = \sigma\left(\boldsymbol{D}_v^{-1/2}\boldsymbol{H}\boldsymbol{W}\boldsymbol{D}_e^{-1}\boldsymbol{H}^\top\boldsymbol{D}_v^{-1/2}\boldsymbol{X}_{pre}\boldsymbol{\Theta}\right), \tag{3}$$

where $\boldsymbol{X}_{pre}$ is the feature after Pre-MLP, $\boldsymbol{\Theta}$ is the filter, and $\sigma(\cdot)$ is the activation. $\boldsymbol{D}_v$ and $\boldsymbol{D}_e$ denote the degree matrices derived from $\boldsymbol{H}$ and weight matrix $\boldsymbol{W}$. As shown in Fig. 2(C), Pre- and Post-MLPs perform channel-wise transformations, while the hypergraph $\mathcal{G}$ is dynamically updated prior to the final residual connection. Multiple HSSH blocks are cascaded to form a Semantic Encoder, which acquires multi-scale high-level topological information.

2.3 Cross-Covariance Semantic Injection (CCSI)

To bridge the semantic gap, the CCSI module (Fig. 2(D)) dynamically rectifies local CNN representations by injecting global topological guidance via a dual-

Table 1. Quantitative comparison of S^{3}H-Net with SOTA methods on five medical imaging modalities. D and H denote DSC (%) and HD95 (mm) respectively. **Best** and second-best results are highlighted. ● CNN, ▲ Trans., ♦ SSM, ■ Graph.

Method & Venues	Colonoscopy				Dermos.		Ultrasound		Micros.		MRI		p-val
	CVC [2]		ETIS [22]		ISIC18 [6]		USG [16]		DSB18 [3]		PROM.[14]		
	D↑	H↓	D↑	H↓	D↑	H↓	D↑	H↓	D↑	H↓	D↑	H↓	
● nnUNet [11] Nat.Meth.'21	90.2	20.1	81.7	36.0	89.2	14.7	77.3	36.3	91.6	4.30	89.4	31.6	4e-5
● EMCAD [18] CVPR'24	92.2	16.1	87.9	32.1	88.9	15.3	80.9	38.8	91.3	3.81	88.1	35.6	2e-4
● MADGNet [15] CVPR'24	91.9	20.5	78.5	39.6	88.8	15.6	82.8	26.0	92.1	3.61	88.6	41.5	2e-4
▲ H2Former [10] TMI'23	77.6	42.1	60.9	78.1	88.7	14.7	72.4	53.9	90.9	3.70	86.7	38.7	8e-6
▲ TransUNet [5] MIA'24	92.1	13.8	69.2	125.	88.0	17.7	65.1	40.3	86.9	5.59	74.3	68.6	9e-6
▲ nnWNet [25] CVPR'25	91.0	14.2	77.0	53.9	89.5	15.4	65.2	59.9	91.7	3.88	89.6	29.6	1e-4
♦ VM-UNet [19] TOMM'24	91.7	16.2	81.7	62.8	89.7	13.6	84.2	21.4	91.8	3.72	86.0	56.9	5e-4
♦ U-RWKV [24] MICCAI'25	93.2	12.4	88.6	20.8	89.3	14.1	80.8	30.3	92.6	3.59	89.6	30.7	1e-3
■ G-CASC [17] WACV'24	91.6	15.5	85.6	34.3	89.3	14.6	81.9	37.8	91.4	3.77	84.9	83.3	6e-5
■ AHGNN [4] MICCAI'24	91.3	21.6	88.5	31.6	90.3	14.1	86.7	14.9	92.4	3.51	89.7	27.9	2e-3
■ S^{3}H-Net (Ours)	**94.7**	**6.49**	**95.4**	**10.8**	**92.0**	**11.9**	**87.5**	**12.3**	**92.7**	**3.15**	**91.3**	**21.3**	-

branch cross-covariance mechanism. We first align the topological node features with the pixel grid to obtain the semantic map $\hat{\boldsymbol{F}}_{hg}$ via the *node inversion* (Sect. 2.1). After that, we extract channel descriptors $\boldsymbol{v}_{hg}, \boldsymbol{v}_p$ from $\hat{\boldsymbol{F}}_{hg}$ and the pixel space features $\boldsymbol{F}_p$ using GAP and FC layers. Note that $\boldsymbol{F}_p$ is pre-processed through a 1×1 convolution layer to align channel dimension. A cross-covariance affinity map $\mathcal{M} \in \mathbb{R}^{C_{hg} \times C_{hg}}$ is then computed as $\mathcal{M} = \text{Softmax}\left(\boldsymbol{v}_p(\boldsymbol{v}_{hg})^\top\right)$. Acting as a semantic bridge, $\mathcal{M}$ reciprocally recalibrates the channel responses, allowing the model to selectively emphasize category-specific features while suppressing semantic ambiguities. The final feature injection is formulated as:

$$\boldsymbol{F}_{inj} = (\hat{\boldsymbol{F}}_{hg} \odot \mathcal{M}\boldsymbol{v}_{hg}) + (\boldsymbol{F}_p \odot \mathcal{M}^\top \boldsymbol{v}_p) + \boldsymbol{F}_p, \tag{4}$$

where $\odot$ denotes element-wise multiplication. This strategy effectively integrates high-order context into local receptive fields, ensuring robust semantic consistency across the target anatomy.

2.4 Loss Function

To facilitate gradient flow and refine features across multiple scales, we employ a deep supervision strategy. The total objective function is a weighted sum of the Dice similarity coefficient ($\mathcal{L}_{Dice}$) and Binary Cross-Entropy ($\mathcal{L}_{BCE}$) computed at each decoder stage s:

$$\mathcal{L}_{total} = \sum_{s=1}^{4} w_s \left(\mathcal{L}_{Dice}(\hat{\boldsymbol{Y}}_s, \boldsymbol{Y}) + \mathcal{L}_{BCE}(\hat{\boldsymbol{Y}}_s, \boldsymbol{Y}) \right), \tag{5}$$

where $\hat{\boldsymbol{Y}}_s$ and $\boldsymbol{Y}$ denote the prediction at scale s and the ground truth. We configure the supervision weights as $w = \{0.125, 0.25, 0.5, 1.0\}$ to prioritize the final output while guiding early layers. This hybrid loss leverages $\mathcal{L}_{Dice}$ to handle class imbalance and $\mathcal{L}_{BCE}$ for pixel-wise classification accuracy.

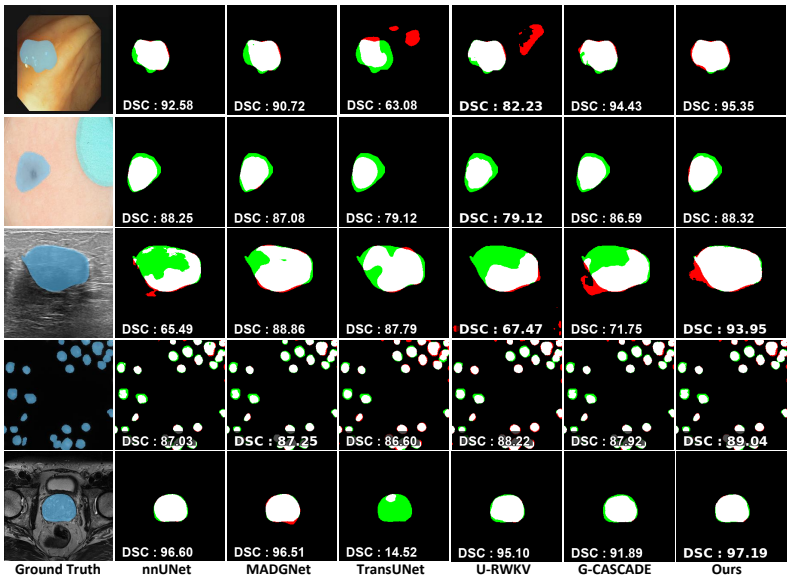

Fig. 4. Qualitative comparison of our S^{3}H-Net with other SOTA methods. Over-segmented areas are marked in red, and under-segmented areas are marked in green.

3 Experiments and Results

3.1 Datasets and Implementation details

Datasets and Metrics. To validate the generalization capability of the proposed S^{3}H-Net, we conduct comprehensive experiments across five distinct medical imaging modalities: Colonoscopy (CVC-ClinicDB [2] and ETIS [22]), Dermoscopy (ISIC2018 [6]), Ultrasound (USG [16]), Microscopy (DSB2018 [3]), and MRI (PROMISE12 [14]). All datasets were split into training, validation, and test sets with a ratio of 8:1:1 except PROMISE12, for which the official challenge split was used. We employ the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95) as the primary evaluation metrics.

Implementation Details. The proposed S^{3}H-Net is implemented using PyTorch. All experiments are conducted on a single NVIDIA 2080Ti GPU (11GB VRAM). We employ the AdamW optimizer with an initial learning rate of 5e-4 and a weight decay of 1e-4. The learning rate is decayed using a cosine annealing scheduler. The batch size is set to 8 and the training process runs for 200 epochs. All input images are resized to 256×256, and data augmentations such as random flipping and rotation are applied during training. We employ EfficientNetv2 [21] as the CNN backbone encoder. Based on our ablation studies (Sect. 3.3), we determine the optimal number of superpixels $M = 512$ and semantic hyperedges $K_{\text{sem}} = 32$. Other hyperparameters are set empirically: SLIC

Table 2. Ablation studies, where all values are reported in DSC (%). (a) Impact of different components. (b) Comparison of graph construction strategies. (c) Hyperparameter sensitivity analysis of K_{sem} and M on CVC-ClinicDB [2] dataset.

(a) Component Effectiveness

Components			Datasets	
HSSH	Sps.	CCSI	ETIS	USG
			92.3	84.4
✓			93.6	85.9
✓	✓		94.8	87.0
✓	✓	✓	**95.4**	**87.5**

(b) Graph Construction

Graph Type	Datasets	
Method	ETIS	USG
ViG [8] (Graph)	93.2	85.6
kNN (Vanilla HG)	93.7	86.2
[4] (Degree-based HG)	94.5	87.1
HSSH (Ours)	**95.4**	**87.5**

(c) Sensitivity Analysis

K_{sem} \ M	128	256	512	1024
16	94.1	94.5	94.3	94.2
32	93.9	94.6	**94.7**	93.8
64	93.4	94.0	94.3	94.1
128	93.2	93.9	93.6	93.3

compactness $\lambda_{xy} = 0.17$, and boundary sensitivity $\lambda_b = 0.5$. Remarkably, S^{3}H-Net incurs moderate computational cost, requiring 54.09M parameters, 10.33G FLOPs, 1106.0 MB peak GPU memory, and 91.17 ms/item inference time, while the differentiable SLIC module contributes only 0.02G FLOPs and 2.91 ms/item, indicating that superpixel node construction is not the main computational bottleneck.

3.2 Comparison with State-of-the-Art Methods

We compare S^{3}H-Net against ten SOTA methods, spanning CNN, Transformer, SSM, and Graph-based architectures, as summarized in Table 1. Quantitative results demonstrate that our method consistently achieves the best performance across all five modalities. Fig. 4 further illustrates the qualitative comparison. It can be observed that S^{3}H-Net generates smoother boundaries and accurately segments small lesions that are missed by competing methods, validating the efficacy of our spatial-semantic hypergraph learning. Statistical t-test analysis ($p < 0.05$) confirms the significance of these improvements.

3.3 Ablation Study

To comprehensively investigate the effectiveness of S^{3}H-Net, we conduct extensive ablation studies, as detailed in Table 2. First, we evaluate the individual components (HSSH, superpixel node, and CCSI), confirming that each module contributes steadily to the overall performance, demonstrating their necessity and complementarity. Second, while the S^{3}H-Net framework can accommodate various graph topologies, comparing it with standard graphs [8], vanilla hypergraph and degree-based hypergraph [4] proves that our proposed HSSH is the optimal structure for capturing complex, high-order correlations. Finally, sensitivity analysis of the hypergraph scale parameters reveals that a moderate configuration ($K_{sem} = 32$, $M = 512$) achieves the best balance, effectively avoiding both insufficient representational capacity and excessive semantic noise.

4 Conclusion

This paper introduces S^{3}H-Net, incorporating a boundary-aware fully differentiable SLIC to leverage morphological superpixels as graph nodes. By construct-

ing the HSSH to model global topology and employing CCSI for adaptive semantic guidance injection, our framework successfully bridges the gap between structured semantics and spatial details. Extensive experiments confirm that S^3H-Net achieves SOTA performance. Although superpixels may still be imperfect under extremely blurred boundaries or noisy appearances, their computational overhead in our framework is minor (0.02G FLOPs and 2.91 ms/item). Ultimately, the role of differentiable superpixels as semantic-aware primitives offers significant general research value, providing a versatile foundation for future explorations in complex anatomical relationship modeling and knowledge-guided clinical decision support.

Acknowledgments. This study was funded by Innovation Platform and Talent Program (2023TP1047).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
3. Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., et al.: Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nature Methods **16**(12), 1247–1253 (2019)
4. Chai, S., Jain, R.K., Mo, S., Liu, J., Yang, Y., Li, Y., Tateyama, T., Lin, L., Chen, Y.W.: A novel adaptive hypergraph neural network forăenhancing medical image segmentation. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 23–33. Springer Nature Switzerland, Cham (2024)
5. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M.P., Zhang, S., Xing, L., Lu, L., Yuille, A., Zhou, Y.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)
6. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 168–172 (2018)
7. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 3558–3565 (2019)

8. Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E.: Vision gnn: An image is worth graph of nodes. In: Advances in Neural Information Processing Systems. vol. 35, pp. 8291–8303 (2022)
9. Han, Y., Wang, P., Kundu, S., Ding, Y., Wang, Z.: Vision hgnn: An image is more than a graph of nodes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19878–19888 (2023)
10. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Transactions on Medical Imaging **42**(9), 2763–2775 (2023)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
12. Jampani, V., Sun, D., Liu, M.Y., Yang, M.H., Kautz, J.: Superpixel sampling networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 363–380. Springer International Publishing, Cham (2018)
13. Li, Y., Zhang, Y., Cui, W., Lei, B., Kuang, X., Zhang, T.: Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation. IEEE Transactions on Medical Imaging **41**(8), 1975–1989 (2022)
14. Litjens, G., Toth, R., van de Ven, W.J.M., Hoeks, C.M.A., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., Strand, R., Malmberg, F., Ou, Y., Davatzikos, C., Kirschner, M., Jung, F., Yuan, J., Qiu, W., Gao, Q., Edwards, P.E., Maan, B., van der Heijden, F., Ghose, S., Mitra, J., Dowling, J., Barratt, D., Huisman, H., Madabhushi, A.: Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. Medical Image Analysis **18**(2), 359–373 (2014)
15. Nam, J.H., Syazwany, N.S., Kim, S.J., Lee, S.C.: Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11480–11491 (2024)
16. Pawłowska, A., et al.: A curated benchmark dataset for ultrasound based breast lesion analysis (Breast-Lesions-USG) (2024), https://doi.org/10.7937/9WKK-Q141
17. Rahman, M.M., Marculescu, R.: G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7713–7722 (2024)
18. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11769–11779 (2024)
19. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. ACM Trans. Multimedia Comput. Commun. Appl. (2025)
20. Suzuki, T.: Hcformer: Unified image segmentation with hierarchical clustering (2023), https://arxiv.org/abs/2205.09949
21. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 10096–10106. PMLR (2021)
22. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017**(1), 4037190 (2017)

23. Yang, Z., Li, C., Ma, J.: Tsldseg: A texture-aware and semantic-enhanced latent diffusion model for medical image segmentation. Pattern Recognition **173**, 112795 (2026)
24. Ye, H., Tang, F., Zhao, P., Huang, Z., Zhao, D., Bian, M., Zhou, h.K.: U-rwkv: Lightweight medical image segmentation withădirection-adaptive rwkv. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. pp. 613–623. Springer Nature Switzerland, Cham (2026)
25. Zhou, Y., Li, L., Lu, L., Xu, M.: nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20852–20862 (2025)