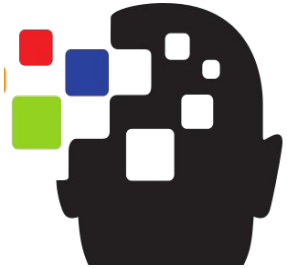


ICML

International Conference

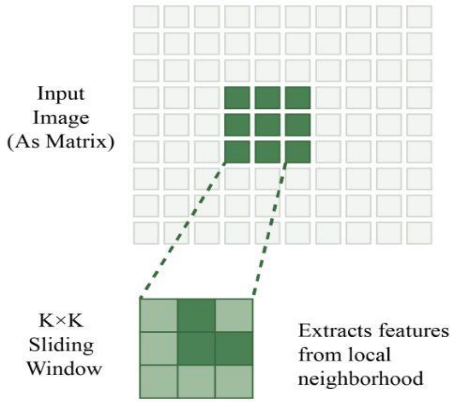
S2M-Net: Spectral-Spatial Mixing with Morphology-Aware Adaptive Loss for Medical Image Segmentation

---

Authors: Sanaullah Chowdhury, Lameya Sabrin

Feature Interaction Paradigms in Medical Image Segmentation

CNN-Based



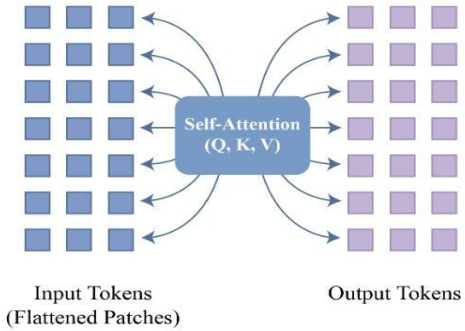
Strengths

- Effective local feature extraction
- Computationally efficient

Limitations

- Limited global context
- Misses long-range dependencies

Transformer-Based



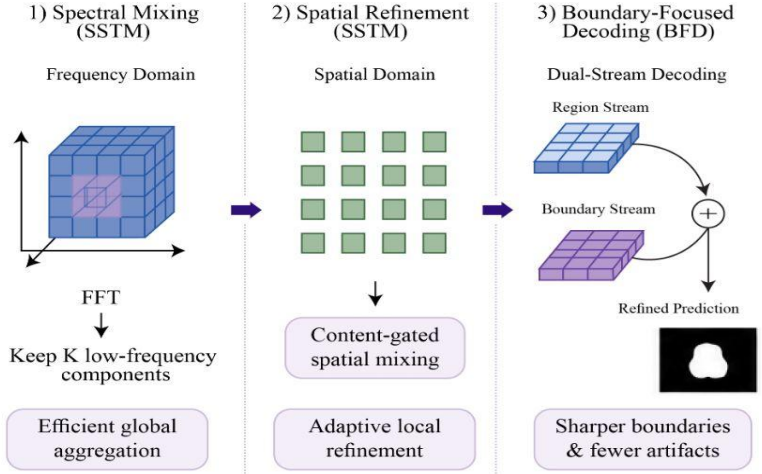
Strengths

- Captures global dependencies
- Powerful long-range context modeling

Limitations

- Quadratic complexity $O(n^2)$
- High memory and computation cost

Our Method (S2M-Net)



Strengths

- Efficient global context via spectral mixing $O(HW\ C \log(HW))$
- Boundary-aware refinement with dual-stream decoding
- Lower complexity with strong performance

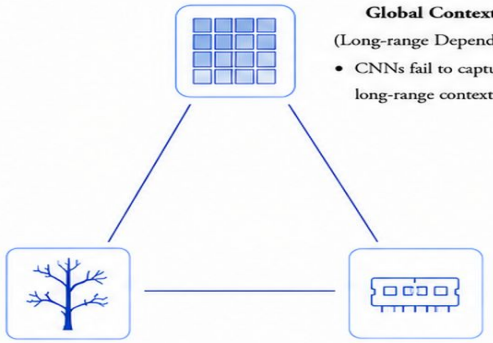
Outcome

Accurate edges, preserved details, and efficient computation

Comparison Aspect	Receptive Field	Context Modeling	Computational Complexity	Memory Usage	Boundary Accuracy	Artifacts
CNN-Based	Local	Local	$O(HW\ C\ K^2)$	Low	Coarse	More
Transformer-Based	Global	Global	$O((HW)^2\ C)$	Very High	High	Moderate
Our Method (S2M-Net)	Global + Local (Adaptive)	Global + Local (Adaptive)	$O(HW\ C \log HW + HW\ C^2)$	Low	High	Fewer

Motivation, Problem Formulation & Contributions

Motivation: The Clinical Trilemma



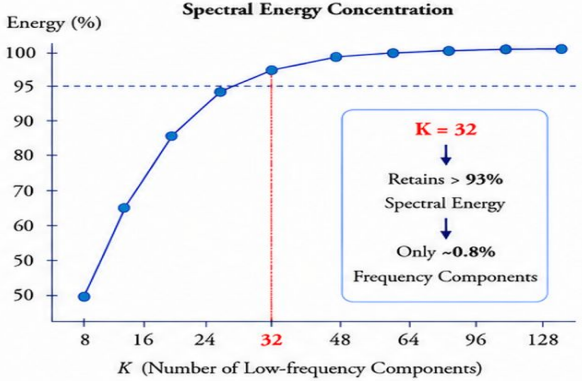
Global Context
(Long-range Dependency)
- CNNs fail to capture long-range context

Boundary Precision
- Thin vessels
- Tumor boundaries
- Surgical instruments

Efficiency
- Transformers are computationally expensive
- $O(n^2)$ attention complexity

Current methods struggle to achieve all three simultaneously.

Core Insight: Spectral Concentration in Medical Images



Key Insight:
Medical images contain highly concentrated low-frequency information. This enables efficient global context modeling without expensive self-attention.

Problem Formulation

We formulate medical image segmentation as:

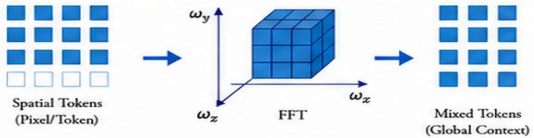
$$(\theta^*, w^*) = \arg\min_{\theta, w} E_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}_{MASL}(y, f_\theta(x); w)\right]$$

Where:
- x : Input image
- y : Ground-truth segmentation
- f_θ : S2M-Net with parameters θ
- w : Adaptive loss weights
- $\mathcal{L}_{MASL}$: Morphology-Aware Adaptive Loss
- $\mathcal{D}$: Data distribution

Goal: Learn optimal network parameters θ and adaptive loss weights w to minimize the MASL.

1) Efficient Global Context Modeling

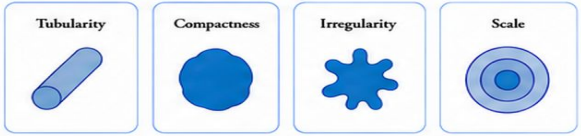
Spectral-Spatial Token Mixing (SSTM)



- FFT-based global aggregation in frequency domain
- $O(HWC\log(HW))$ complexity (vs. $O(n^2)$ attention)
- Captures long-range dependencies efficiently

2) Adaptive Morphology-Aware Supervision

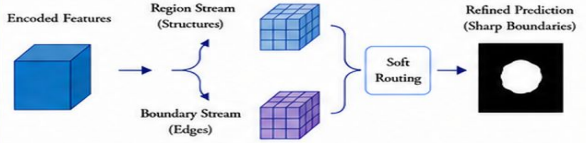
Morphology-Aware Adaptive Loss (MASL)



- Captures key morphological properties of medical structures
- Automatically balances multiple objectives via learnable weights
- Improves generalization across datasets and structures

3) Boundary-Focused Decoding

Boundary-Focused Decoding (BFD)



- Dual-stream decoding with soft routing mechanism
- Focuses on boundaries without extra annotations
- Produces sharper boundaries with fewer artifacts

Architecture Overview

MRF-SE Block (Local Feature Extraction):

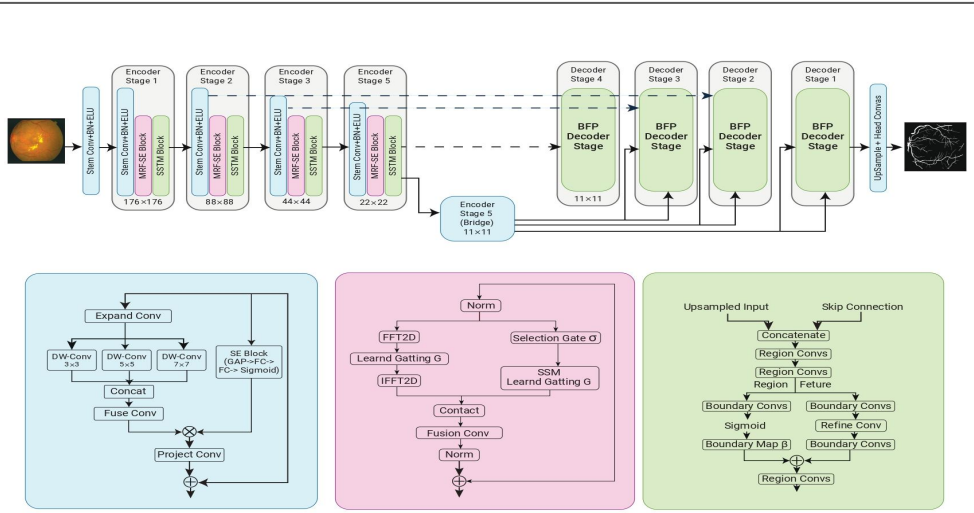
- Parallel DW-Conv with k = {3, 5, 7} captures multi-scale local patterns.
- SE recalibration: GAP → FC → Sigmoid for channel attention.
- Handles both tiny polyps (5 px) and large vessels (>20 px) simultaneously.
- Removing MRF-SE: −5.90% average Dice across 4 datasets.

SSTM Block (Efficient Global Context):

- 2D FFT → retain only K = 32 low-frequency components, just 0.8% of the spectrum.
- Captures >93% spectral energy across all 8 modalities.
- Sub-quadratic complexity: $O(HWC \log(HW) + HWC^2)$ vs. $O((HW)^2C)$ self-attention.
- Content-gated spatial branch with bottleneck d = 16.
- Removing SSTM: −6.70% average Dice across 4 datasets

BFD Decoder (Boundary Specialization):

- Dual-stream: region path for smooth interiors + boundary path for edge sharpness.
- Soft routing: $z = r \odot (1 - \beta) + b \odot \beta$.
- β is learned via backpropagation from MASL, with no edge annotations needed.
- Removing BFD: −5.06% average Dice across 4 datasets.



Morphology-Aware Adaptive Loss

- Automatically weights five objectives per sample: region, boundary, structure, focal, and texture.
- Uses four geometric descriptors: tubularity τ, compactness c, irregularity ι, and size s.
- Generalizes across 15 datasets using a single loss formulation, without manual tuning.
- Learned weights $w_i \in [0.1, 10]$ adapt to target structures, e.g., polyps → boundary, vessels → structure.

Results

Table 1. Architectural comparison under a unified training pipeline. All methods are retrained with identical optimization and preprocessing; MASL is applied to all to isolate architectural effects under a common supervision signal. For transparency, Table 2 evaluates S2M-Net and top baselines under standard Dice+CE loss to confirm that architectural gains are independent of MASL. Mean Dice (%) ± std over 5 runs. †: statistically significant ($p < 0.0033$, Bonferroni-corrected). S2M-Net uses 4.7M parameters (12.8× fewer than TransUNet). FNet and GFNet are excluded as they lack decoders and skip connections required for dense prediction; see Section 2 for discussion. RAPUNet(Lee & Yoo, 2024) results follow the DuckNet data split; see footnote.‡

Dataset	U-Net	U-Net++	PraNet	Swin-Unet	TransUNet	UMamba	DuckNet	RAPUNet	S2M-Net	Model Params
Polyp Segmentation (Endoscopy)										
Kvasir-SEG	90.64±0.41	91.81±0.53	93.03±0.38	93.22±0.44	93.75±0.36	92.40±0.47	94.11±0.33	93.12±0.45	**96.05**±0.28†	4.7M
CVC-ClinicDB	92.18±0.38	92.70±0.42	92.65±0.44	92.52±0.61	87.28±0.52	90.09±0.48	93.89±0.54	95.10±0.38	**95.65**±0.31†	4.7M
CVC-ColonDB	88.12±0.56	88.98±0.49	**92.78**±0.37	91.45±0.43	90.33±0.47	86.62±0.58	92.69±0.39	91.85±0.52‡	90.69±0.52	4.7M
ETIS-LaribDB	91.22±0.44	89.80±0.51	93.46±0.35	94.10±0.33	94.43±0.31	91.23±0.46	94.98±0.29	94.23±0.41	**96.12**±0.30	4.7M
Dermoscopy (Skin Lesions)										
ISIC-2018	86.93±0.48	87.51±0.45	89.05±0.39	89.43±0.42	87.12±0.47	89.23±0.38	90.10±0.43	89.45±0.31	**91.09**±0.36	4.7M
PH2	88.34±0.52	89.83±0.46	91.52±0.38	92.66±0.35	93.56±0.39	90.12±0.44	90.34±0.43	92.15±0.49	**96.67**±0.32†	4.7M
Histopathology & Ultrasound										
GlaS	91.57±0.39	91.40±0.41	90.40±0.46	89.01±0.53	89.92±0.48	91.18±0.42	88.73±0.55	91.67±0.07	**93.83**±0.38	4.7M
BUSI	76.77±0.61	78.35±0.56	79.48±0.52	82.52±0.49	82.11±0.51	79.04±0.58	78.13±0.63	81.34±0.73	**85.07**±0.45†	4.7M
Surgical Robotics (EndoVis-2017)										
Binary	95.93±0.26	95.32±0.29	95.24±0.30	95.24±0.30	95.45±0.28	95.53±0.27	95.11±0.31	95.78±0.36	**95.95**±0.28	4.7M
Multiclass	61.07±1.12	70.93±0.89	73.86±0.76	73.22±0.78	74.74±0.72	74.29±0.75	73.88±0.87	74.81±0.19	**83.43**±0.63†	4.7M
Brain Tumor MRI										
BraTS2020	65.67±0.84	66.05±0.82	66.44±0.81	69.86±0.73	68.36±0.77	62.88±0.91	66.93±0.79	67.82±0.12	**79.96**±0.58†	4.7M
Cardiac MRI										
ACDC	89.52±0.42	90.15±0.45	91.11±0.38	92.32±0.37	92.44±0.36	91.57±0.48	92.13±0.38	91.45±0.20	**93.09**±0.43	4.7M
Retinal Vessel Segmentation										
STARE	81.18±0.54	81.73±0.51	82.94±0.47	83.86±0.44	83.10±0.46	82.05±0.58	81.41±0.62	82.34±0.38	**84.45**±0.48	4.7M
DRIVE	79.02±0.56	80.18±0.59	81.94±0.49	83.07±0.48	81.45±0.50	82.18±0.58	81.39±0.57	82.94±0.12	**84.06**±0.49	4.7M
CHASE-DB	81.51±0.52	81.65±0.51	81.47±0.52	83.12±0.49	82.93±0.47	82.90±0.59	83.16±0.58	83.21±0.29	**84.95**±0.48	4.7M
Average	83.98	85.09	86.29	87.04	86.46	85.42	86.47	86.72	**90.07**	**4.7M**

‡RAPUNet follows PraNet data split for polyp datasets; S2M-Net uses DuckNet split. Direct numerical comparison on CVC-ColonDB is not fully appropriate without matching distributions. FNet/GFNet excluded: classification-only, no decoder. (Lee & Yoo, 2024; Lee-Thorp et al., 2022; Rao et al., 2021)

Results and Ablation Studies

Table 2. Architecture vs. MASL disentanglement. S2M-Net evaluated under standard Dice+CE loss to isolate architectural gains from loss design. Dice (%).

Method	Loss	Kvasir	DRIVE	EV17-MC	BraTS20	Avg
TransUNet	Dice+CE	92.17	80.92	71.83	66.42	77.84
Swin-Unet	Dice+CE	91.84	82.14	70.54	67.91	78.11
S2M-Net	Dice+CE	93.89	82.74	78.91	74.83	**82.59**
S2M-Net	MASL	96.05	84.06	83.43	79.96	**85.88**
Arch. gain (S2M vs. best@Dice+CE)		+1.72	+0.60	+7.08	+6.92	**+4.48**
MASL gain (MASL vs. Dice+CE)		+2.16	+1.32	+4.52	+5.13	**+3.29**

Table 4. Spectral energy retention across medical imaging modalities. $K = 32$ retains 94.8% average energy while representing only 0.83% of frequency coefficients at 352×352 resolution. Medical images show stronger concentration than natural images, validating domain-specific truncation.

Modality	K=16	K=24	K=32	K=48
Colonoscopy	89.3±2.1%	93.7±1.4%	95.2±1.2%	97.1±0.9%
Fundus	91.2±1.8%	94.8±1.2%	96.4±0.9%	97.8±0.7%
Brain/Cardiac MRI	87.6±2.4%	92.1±1.6%	94.1±1.3%	96.2±1.0%
Ultrasound	85.9±3.2%	90.8±2.1%	93.3±1.7%	95.4±1.3%
Average	88.5±2.4%	92.9±1.6%	94.8±1.3%	96.6±1.0%

Table 5. Per-structure segmentation performance comparing $K = 32$ truncation to full-resolution FFT. Mean Dice degradation is 0.24% across tested structures, with differences within confidence intervals. Thin structures (< 3px vessels, instrument tips) show degradation $< 0.5\%$.

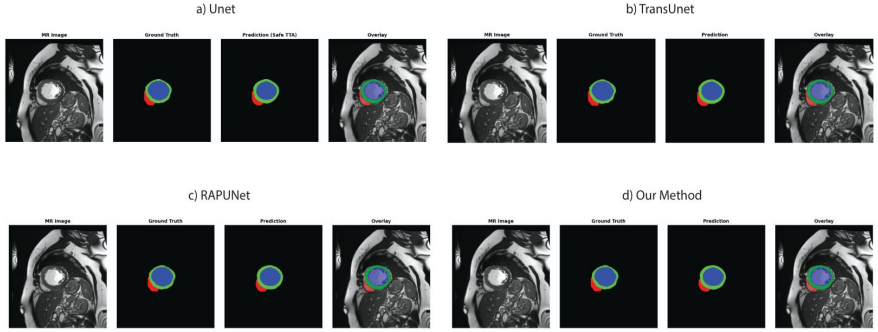
Dataset	Structure	K=32	Full	Δ
DRIVE	Thin vessels (<3px)	83.6±0.8	83.9±0.9	-0.3%
	Medium vessels	86.7±0.5	86.9±0.6	-0.2%
Kvasir	Small polyps	95.1±0.9	95.3±1.0	-0.2%
	Medium polyps	96.9±0.5	97.0±0.6	-0.1%
EndoVis17	Instrument tips	82.3±1.4	82.7±1.5	-0.4%

Table 3. Ablation study on four medical imaging datasets.Dice coefficient (%) reported for systematic component removal. Full model (bold) includes all components. Differences ≥1.0% are statistically significant (p<0.001).

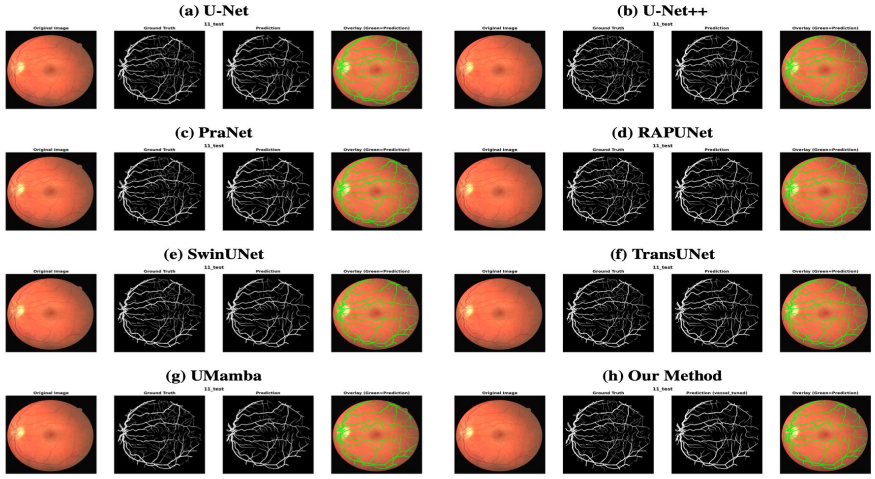
Configuration	Kvasir	DRIVE	GlaS	PH2
Full Model (Ours)	**96.05**	**84.06**	**93.83**	**96.67**
MASL Loss Components				
w/o Core Loss	92.31	81.12	77.32	93.82
w/o Boundary Loss	87.96	80.45	90.85	89.91
w/o Structure Loss	94.42	82.67	90.17	93.88
w/o Scale-Aware Focal	95.15	83.21	90.43	94.56
w/o Texture Loss	95.68	83.04	88.67	94.01
Adaptive Mechanisms				
Fixed Weights	88.92	78.34	90.38	91.45
No Morphology Modulation	89.37	79.78	90.28	91.92
SSTM Frequency Truncation				
K=16	89.87	80.92	88.57	92.34
K=24	93.56	82.45	90.43	93.98
K=32	96.05	84.06	93.83	96.67
K=48	95.98	83.71	89.89	93.38
K=64	95.75	83.52	90.42	93.11
Architecture Modules				
w/o SSTM	88.45	77.78	86.92	90.67
w/o BFD Decoder	89.87	79.89	88.34	92.28
w/o MRF-SE	89.23	78.45	87.56	91.78
Vanilla U-Net Baseline	85.67	74.34	86.48	88.45

Qualitative Studies

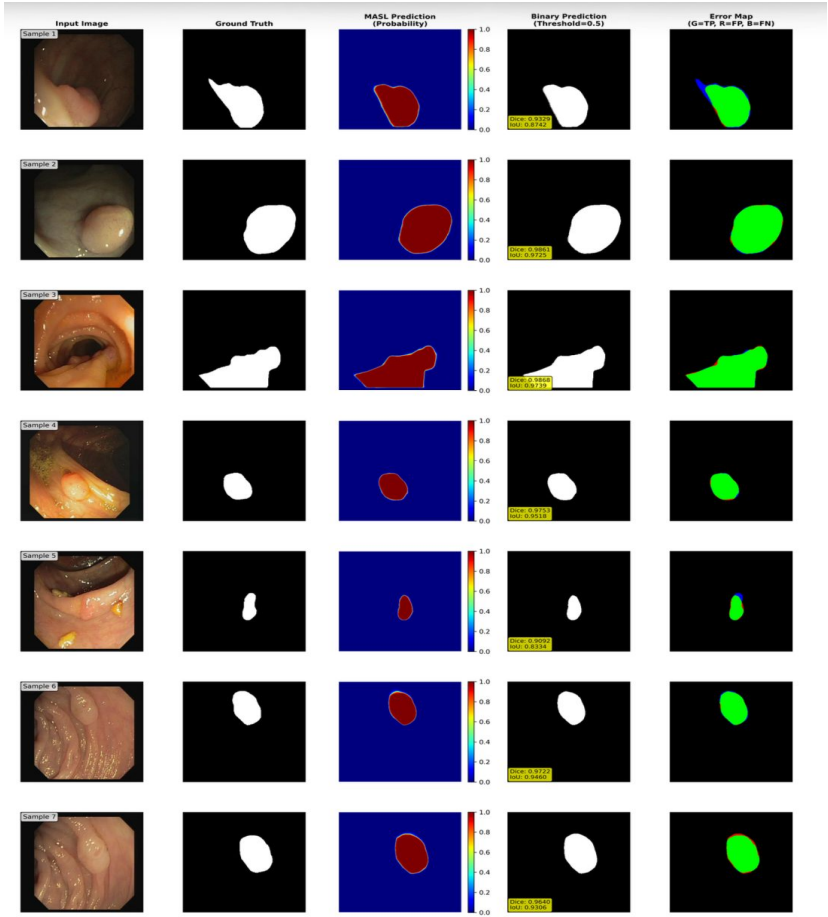
ACDC Multiclass Segmentation



DRIVE Blood Vessel Segmentation



KVASIR-SEG Polyp Segmentation



Thank you

Authors: Sanaullah Chowdhury, Lameya Sabrin sanaullahashfat@gmail.com, slameya.work@gmail.com