ROLE-PRESERVING MULTISCALE EVIDENCE LEARNING FOR COMPACT MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Compact medical image segmentation requires object-level discrimination without sacrificing low-contrast or irregular boundaries. Many compact encoder–decoder networks merge coarse semantics, regional structure, and fine texture into a single stream; an erroneous coarse response can therefore become a coherent false-positive object, whereas strong filtering can remove a genuine target. We propose a role-preserving multiscale evidence framework that assigns object evidence construction, regional continuity, boundary reconstruction, and class readout to explicit paths. At $H/16$, local and regional observations are related before a compact KAN mapper modulates a stable high-dimensional convolutional carrier. At higher resolutions, semantic, regional, and detail representations retain separate carriers and interact through geometric correspondence during transport. A cross-scale prototype head then allocates dense evidence to image-conditioned foreground and background prototypes instead of drawing an independent mask. Under a common three-seed protocol on three public benchmarks spanning ultrasound, endoscopy, and histopathology, our 5.79M-parameter model obtains mean IoUs of 73.97%, 85.06%, and 82.78% on BUSI, CVC, and GlaS, respectively. It achieves the highest BUSI and CVC mean IoU among the evaluated methods and improves the three-dataset average IoU from 79.52% for U-KAN to 80.60%. These results support role-explicit evidence transport as an effective architectural choice under the evaluated compact-segmentation protocol.

1 INTRODUCTION

Medical image segmentation must identify complete objects, recover weak or irregular boundaries, and limit parameter growth. These goals can conflict: stronger local restoration may turn background structures into false-positive objects, whereas aggressive semantic filtering may remove low-contrast targets. The central problem is therefore to determine which evidence should establish object identity, maintain regional continuity, or refine the boundary.

U-shaped networks address scale variation through progressive decoding and skip connections (Ronneberger et al., 2015); attention gates and nested pathways regulate this transfer (Oktay et al., 2018; Zhou et al., 2020). Yet compact decoders commonly merge coarse semantics, regional structure, and fine texture into one stream. Later layers then cannot tell whether a response is reliable object evidence or merely plausible local appearance, and additional fusion may increase cross-scale interference.

Kolmogorov–Arnold Networks (KANs) offer expressive learnable univariate mappings (Liu et al., 2025), and U-KAN applies them to tokenized intermediate representations (Li et al., 2025). We instead use compact KAN modules to model relations among complementary observations and regulate a stable high-dimensional convolutional carrier, making the nonlinear path spatially and functionally explicit.

We therefore propose *role-preserving multiscale evidence learning*. A role denotes an information source's functional responsibility and permitted operation, not its resolution alone. Preservation does not prohibit interaction; it delays irreversible mixing until role-specific transformations have been formed. The architecture implements this rule in three ways: relations modulate without replacing a carrier, representations share geometric correspondence without sharing values, and context allocates prototype evidence without drawing an independent mask.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our contributions are summarized as follows:

- We introduce an end-to-end compact U-shaped architecture organized by an explicit role-preserving constraint: object, regional, and detail carriers may exchange conditions, but no common fused tensor replaces them before their designated decoding operations.
- We instantiate this constraint through local proposal and regional verification, carrier–relation factorization, role-specific cross-scale transport, and an image-conditioned multi-prototype readout. Compact KAN mappings regulate low-dimensional evidence relations rather than replacing the full convolutional carrier.
- Across three datasets and three seeds, the resulting 5.79M-parameter model attains the best average IoU and F1 among the evaluated methods, with the highest mean IoU on BUSI and CVC.

2 Related Work

Segmentation and role separation. U-Net (Ronneberger et al., 2015), Attention U-Net (Oktay et al., 2018), UNet++ (Zhou et al., 2020), and nnU-Net (Isensee et al., 2021) establish encoder–decoder baselines; transformers add global interaction (Chen et al., 2024; Hatamizadeh et al., 2022; Cao et al., 2022; Tang et al., 2022), while MedNeXt modernizes convolutional processing (Roy et al., 2023). Parameter-compact designs include UNeXt (Valanarasu & Patel, 2022), EGE-UNet (Ruan et al., 2023), Rolling-Unet (Liu et al., 2024), EMCAD (Rahman et al., 2024), and HCI (Cheng et al., 2025). Region–boundary separation is explored by BiSeNet, Gated-SCNN, Co-Net, and boundary-aware objectives (Yu et al., 2018; Takikawa et al., 2019; Liu et al., 2022; Kervadec et al., 2021). We instead study an information-organization constraint spanning object, regional, detail, and allocation roles.

Relation, transport, and allocation. KAN uses learnable edge-wise univariate functions (Liu et al., 2025), and U-KAN applies it to tokenized features (Li et al., 2025); we confine it to compressed evidence relations that modulate, but do not replace, a convolutional carrier. Content-aware upsamplers learn reassembly kernels, affiliations, or sampling locations (Wang et al., 2019; Lu et al., 2022a;b; Liu et al., 2023). Our transport instead shares geometric correspondence across roles while retaining independent value projections. Unlike OCR and ProtoSeg (Yuan et al., 2020; Zhou et al., 2022), our final allocator lets context reweight foreground/background evidence but not emit an independent mask.

3 Method

3.1 Principle and Overview

Let $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ be an input image and $\mathbf{Y} \in \{0,1\}^{H\times W}$ its binary mask. The network predicts a logit map $\mathbf{Z}$ and a probability map $\mathbf{P} = \sigma(\mathbf{Z})$. Our design follows a role-preserving principle: features responsible for object verification, regional continuity, and local detail may exchange evidence, but they are not forced into a single representation before their respective decisions are made.

Operational definition. For role i, let $\mathbf{F}_i$ denote its carrier and d_i its designated operation. Given multiple roles, we call an exchange role-preserving when

$$\mathbf{r} = \mathcal{G}(\mathcal{P}_1(\mathbf{F}_1), \ldots, \mathcal{P}_m(\mathbf{F}_m)), \qquad \widetilde{\mathbf{F}}_i = \mathbf{F}_i + \mathcal{A}_i(\mathbf{r}) \odot \Delta_i(\mathbf{F}_i), \tag{1}$$

where the shared relation $\mathbf{r}$ conditions a role-specific residual, while each carrier remains available until d_i. Thus, cross-role information may modulate a carrier but cannot replace all carriers with one early fused representation. This is an architectural constraint, not a claim of statistical independence. In our network, the designated operations are object verification at $H/16$, regional reconstruction at $H/8$, boundary and detail reconstruction at $H/4$–H, and class allocation in prototype space. In short: relate without replacing, transport without overwriting, and allocate without redrawing.

Figure 1 gives the complete information flow. The three principal stage diagrams are placed beside their corresponding method descriptions below; additional implementation views remain in the appendix.

Table 1: One role-preserving constraint instantiated at three decoding stages. A relation may condition a carrier, but does not replace it before the carrier's designated operation.

Stage	Preserved carrier(s)	Permitted cross-role action
$H/16$	Convolutional object carrier	Bounded relation-conditioned residual for object verification
$H/8$–H	Semantic, regional, and detail carriers	Shared geometric correspondence with independent transported values
Readout	Pixel carrier and class prototypes	Bounded prototype-mixture allocation without an independent mask

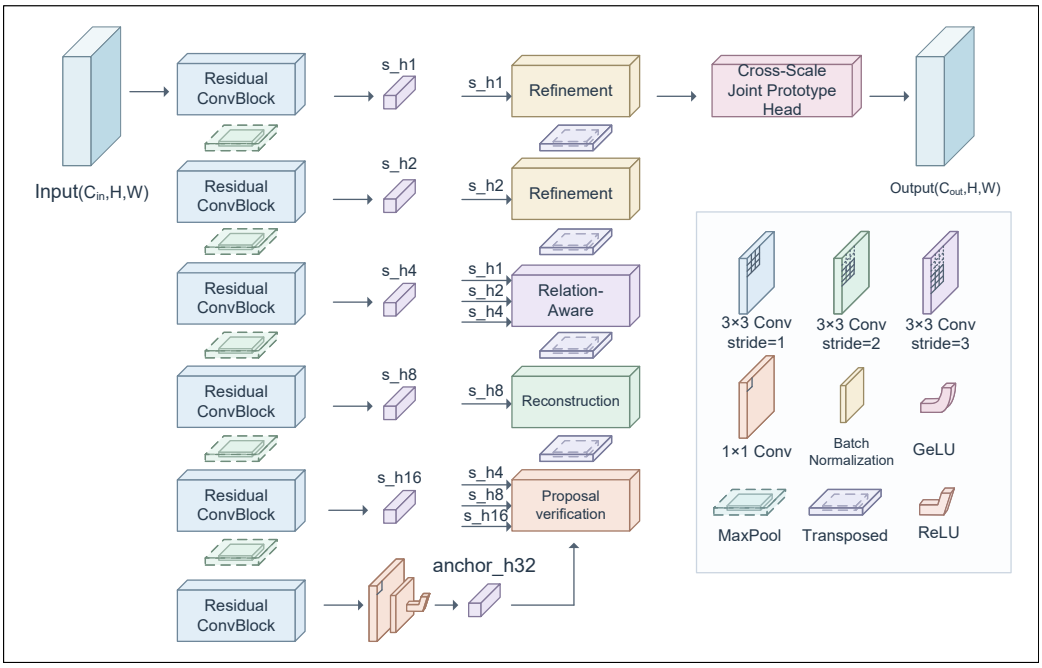

Figure 1: Overview of the proposed role-preserving multiscale evidence learning framework. The encoder supplies representations at multiple resolutions, while the decoder keeps object evidence, regional continuity, and detail reconstruction distinct until their designated readout stages.

3.2 Compact Residual Encoder and Minimal Neck

The encoder contains residual convolutional stages (He et al., 2016) at resolutions $\{H, H/2, H/4, H/8, H/16, H/32\}$. We denote the five skip features by $\mathbf{S}_r$, $r \in \{1, 2, 4, 8, 16\}$, and the bottleneck by $\mathbf{B}_{32}$. Each stage applies two 3×3 convolutions and a residual shortcut,

$$\mathbf{S}_r = \rho\big(\mathcal{F}_r(\mathrm{Pool}(\mathbf{S}_{r/2})) + \mathcal{Q}_r(\mathrm{Pool}(\mathbf{S}_{r/2}))\big)\,, \tag{2}$$

where ρ is ReLU and $\mathcal{Q}_r$ is an identity or 1×1 projection. Same-scale residual processing precedes every downsampling step, preserving a direct path for local appearance while progressively increasing the receptive field. Unlike the other skip levels, $\mathbf{S}_1$ is also retained as a full-resolution detail path for the final reconstruction. A single 1×1 convolution forms the semantic anchor $\mathbf{A}_{32}$ from $\mathbf{B}_{32}$; no attention or token block is placed in the neck. The decoder then produces $\mathbf{X}_{16}, \mathbf{X}_8, \mathbf{X}_4, \mathbf{X}_2, \mathbf{X}_1$. Channels are $[16, 32, 64, 128, 256, 256]$ in the encoder and $[128, 96, 64, 32, 16]$ in the decoder. Detailed stage diagrams are provided in Appendix A.

3.3 Local Proposal and Regional Verification

The $H/16$ stage establishes object evidence before high-resolution detail is introduced. Upsampling $\mathbf{A}_{32}$ gives $\mathbf{U}_{16}$. We form a recall template from $\mathbf{S}_{16}$, a conservative template from $\mathbf{U}_{16}$, and a candidate carrier from both:

$$\mathbf{T}^{\mathrm{rec}} = \phi_r(\mathbf{S}_{16}), \quad \mathbf{T}^{\mathrm{safe}} = \phi_s(\mathbf{U}_{16}), \quad \mathbf{C} = \phi_c([\mathbf{U}_{16}, \mathbf{S}_{16}]). \tag{3}$$

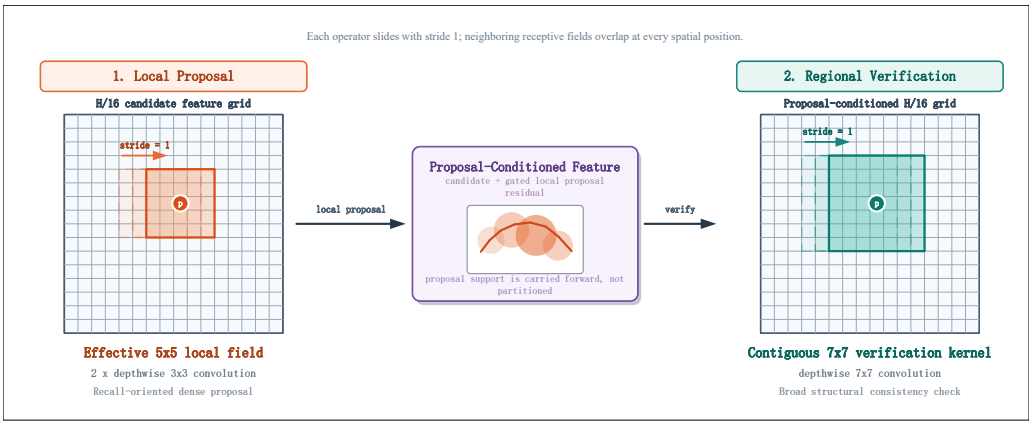

Figure 2: The proposed $H/16$ object-evidence stage. Local proposal and regional verification are retained as distinct evidence sources before compact relation mapping modulates the convolutional carrier.

Cross-scale projections additionally provide structural evidence $\mathbf{E}^{\mathrm{str}}$ from $\mathbf{S}_4, \mathbf{S}_8$ and semantic evidence $\mathbf{E}^{\mathrm{sem}}$ from $\mathbf{S}_{16}, \mathbf{U}_{16}$. These are read-only conditions rather than replacement features.

The local branch applies two overlapping depthwise 3×3 units with stride one, giving an effective 5×5 field at every location. It returns a recall-oriented proposal feature $\mathbf{L}$ and support p_l. The proposal is written into the candidate through a bounded support-dependent residual before the regional observation in our implementation:

$$\mathbf{C}^p = \mathbf{C} + \mathcal{A}_p(p_l) \odot \Delta_p(\mathbf{C}, \mathbf{L}). \tag{4}$$

A contiguous depthwise 7×7 unit then evaluates $\mathbf{C}^p$ and produces regional evidence $\mathbf{V}$ and support p_v. The method retains both observations and their agreement and conflict; Appendix E shows that replacing this implementation order with parallel observations produces comparable average accuracy. We therefore regard the ordering as an implementation choice, rather than as an independently established source of improvement.

We explicitly retain agreement $a = p_l p_v$ and conflict $c = p_l(1-p_v)$. A reliability estimate q controls nonnegative bounded agreement and conflict operators $\mathcal{B}_+$ and $\mathcal{B}_-$:

$$\widetilde{p} = \mathrm{clip}_{[\epsilon, 1-\epsilon]}(p_l + \mathcal{B}_+(a, q, p_l) - \mathcal{B}_-(c, q, p_l)), \tag{5}$$

$$p_s = \sigma(\mathrm{logit}(\widetilde{p}) + \mathcal{K}_s([p_l, p_v, a, c])), \tag{6}$$

where $\mathcal{K}_s$ is a compact KAN relation mapper. The bounded operators limit write authority; their exact parameterization is reported in Appendix B and is fixed across datasets. A post-fusion reader combines p_s, the region context, and the convolutional carrier to obtain $\mathrm{support}_{16}$. Regional verification is also preserved as an independent feature $\mathbf{V}_{16}$ for later semantic conditioning.

3.4 Compact Nonlinear Relation Mapping

KAN is used to model relations, not to replace the full spatial carrier. This placement is deliberate: the relation coordinates encode heterogeneous agreement and conflict terms at low channel dimension, for which learnable edge-wise univariate mappings provide nonlinear calibration without applying a KAN transform to every carrier channel. We do not assume that KAN is universally superior to an MLP; Section 4.3 tests this choice within the same architecture using approximately parameter-matched mappings. A two-layer convolutional branch first produces Δ_{cnn} from $\mathbf{C}$. Local support modulates this update, after which the stable 128-channel carrier is

$$\bar{\mathbf{X}}_{16} = \mathbf{C} + (1 + \alpha_m\, m(p_l)) \odot \Delta_{\mathrm{cnn}}. \tag{7}$$

We project the carrier, proposal, verification, structural evidence, and semantic evidence to 32 channels. For two projected tensors $\mathbf{A}, \mathbf{B}$, their relation is represented as

$$\mathcal{R}(\mathbf{A}, \mathbf{B}) = \phi\big([\mathbf{A}, \mathbf{B}, \mathbf{A} \odot \mathbf{B}, |\mathbf{A} - \mathbf{B}|]\big). \tag{8}$$

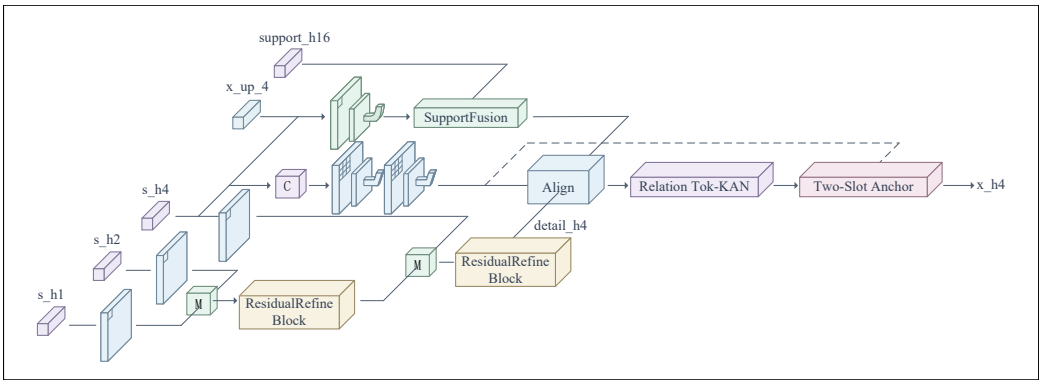

Figure 3: The $H/4$ reconstruction stage. Regional continuity and boundary detail are kept as separate roles and combined through a conditioned residual update before finer-resolution transport.

A local template is obtained from proposal–structure relations. A regional template is constructed by first relating verification to semantic evidence and then relating that result to the local template. Their compressed joint relation $\mathbf{R}_{16}$ is mapped by

$$\widehat{\mathbf{R}}_{16} = \mathrm{KAN}_2(\mathrm{DWConv}_{3\times3}(\mathrm{KAN}_1(\mathrm{LN}(\mathbf{R}_{16})))) \,. \tag{9}$$

Two affine fields $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ are predicted from $\widehat{\mathbf{R}}_{16}$ and converted into a bounded residual

$$\Delta_{\mathrm{kan}} = \bar{\mathbf{X}}_{16} \odot b_\gamma(\boldsymbol{\gamma}) + b_\beta(\boldsymbol{\beta}), \quad \mathbf{X}_{16} = \phi_o(\bar{\mathbf{X}}_{16} + \Delta_{\mathrm{kan}}). \tag{10}$$

Thus the low-dimensional nonlinear path relates evidence without replacing the high-dimensional carrier, which remains directly available to the decoder.

3.5 Role-Preserving Cross-Scale Reconstruction

At $H/8$, the upsampled $\mathbf{X}_{16}$ retains body semantics while $\mathbf{S}_8$ serves as a spatial candidate. Rather than concatenating them, their agreement and conflict condition a residual reconstruction of regional continuity. Appendix A.1 gives the exact relation tensor and multibranch implementation.

The initial $H/4$ carrier $\mathbf{X}_4^0$ is produced by conventional upsampling with $\mathbf{S}_4$. We then form a detail template $\mathbf{D}_4$ by progressively aligning $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_4$ using fixed anti-aliased downsampling and residual refinement. A semantic template $\mathbf{Q}_4$ is obtained from $\mathbf{X}_8$ and softly conditioned by $\mathrm{support}_{16}$. Region and boundary relations are kept separate:

$$\mathbf{R}_4^{\mathrm{reg}} = \phi_r([\mathbf{X}_4^0, \mathbf{Q}_4, \mathbf{D}_4 \odot \mathbf{Q}_4]), \tag{11}$$

$$\mathbf{R}_4^{\mathrm{bnd}} = \phi_b([\mathbf{X}_4^0, \mathbf{D}_4, |\mathbf{D}_4 - \mathbf{Q}_4|]). \tag{12}$$

Their interaction is mapped in relation space and conditioned by the $H/16$ verification template. A permutation-equivariant two-slot anchor avoids assuming in advance which latent slot represents the target. The selected anchor corrects the relation, yielding

$$\mathbf{X}_4 = \mathbf{X}_4^0 + \Delta_4, \qquad \Delta_4 = \phi_\Delta(\mathbf{R}_4^{\mathrm{cond}} + \Delta_4^{\mathrm{anchor}}). \tag{13}$$

Figure 3 summarizes these responsibilities.

Before the $H/4 \to H/2$ transition, we expose three carriers: $\mathbf{X}_4^0$ (common), the conditioned semantic relation, and the detail template. A correspondence module predicts one geometry field from $\mathbf{X}_4$ and $\mathbf{S}_2$. For each of four subpixel phases it assigns a distribution over a 3×3 low-resolution neighborhood,

$$w_{p,n} = \mathrm{softmax}_n\left(\frac{\langle \mathbf{q}_p, \mathbf{k}_n \rangle}{\sqrt{d}} + b_{p,n}\right), \quad p \in \{1, \ldots, 4\},\ n \in \mathcal{N}_9. \tag{14}$$

The same w transports all roles, enforcing geometric consistency, while independent value projections prevent their contents from overwriting one another: roles share where, not what. A spatial softmax selects among them only after $\mathbf{S}_2$ becomes available. At $H/1$, the full-resolution skip $\mathbf{S}_1$ restores image-aligned texture; the $H/4$ anchor is read only to select a local residual and is never recursively added as a carrier.

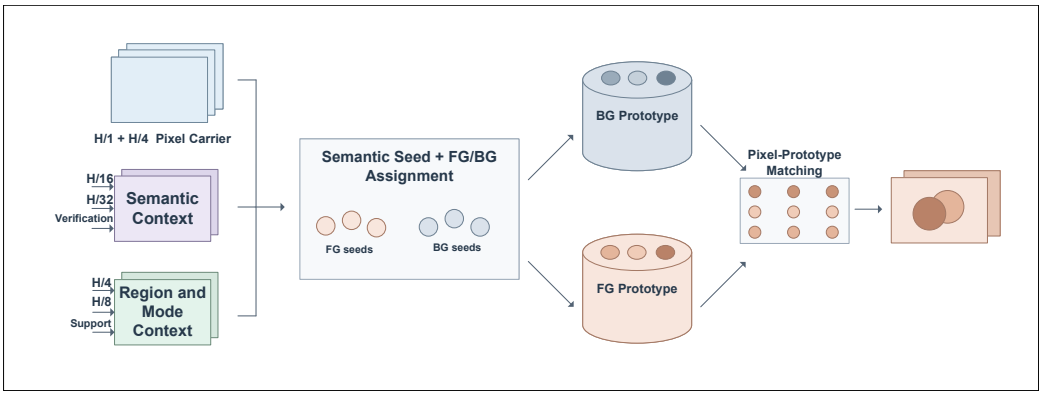

Figure 4: Prototype-based cross-scale readout. Pixel evidence is matched to image-conditioned foreground and background prototypes, while multiscale context allocates mixture weights instead of emitting an unconstrained additive mask.

3.6 Cross-Scale Prototype Readout

The readout in Figure 4 separates pixel evidence from class allocation. A 32-channel pixel carrier is constructed from $\mathbf{X}_1$ and upsampled $\mathbf{X}_4$ and normalized as $\mathbf{F}$. Semantic context combines $\mathbf{X}_{16}$, $\mathbf{V}_{16}$, and $\mathbf{A}_{32}$; regional context combines this semantic context with $\mathbf{X}_8$; and $\mathbf{X}_4$ provides local mode evidence. These contexts predict a semantic seed s and two residual modes for each class. Together with one global mode, they define three foreground and three background weight maps w_k^+ and w_k^-. This fixed three-mode implementation represents one global mode plus two residual modes; A7 tests multiple modes against a single prototype pair, but we do not claim that three is a universally optimal count.

Image-conditioned prototypes are weighted averages of the pixel carrier,

$$\mathbf{p}_k^{\pm} = \frac{\sum_i w_{k,i}^{\pm} \mathbf{F}_i}{\sum_i w_{k,i}^{\pm} + \epsilon}, \qquad d_{k,i}^{\pm} = \langle \widehat{\mathbf{F}}_i, \widehat{\mathbf{p}}_k^{\pm} \rangle. \tag{15}$$

The global foreground prototype additionally receives the pooled semantic context. Foreground and background prototypes are jointly disentangled by subtracting their soft overlap context and renormalizing. Spatial mode priors and prototype separation determine initial mixtures $\pi_k^{\pm}$. Dense cross-scale context may only correct their log-mixture weights through a bounded function b_δ,

$$\widetilde{\pi}_{k,i}^{\pm} = \mathrm{softmax}_k \left(\log(\pi_{k,i}^{\pm} + \epsilon) + b_\delta(\delta_{k,i}^{\pm}) \right). \tag{16}$$

It cannot emit an independent segmentation logit. Class evidence is aggregated with a temperature-controlled log-sum-exp,

$$e_i^{\pm} = \tau \log \sum_k \exp \left(d_{k,i}^{\pm}/\tau + \log \widetilde{\pi}_{k,i}^{\pm} \right), \qquad Z_i = \eta(e_i^+ - e_i^-) + b. \tag{17}$$

This bounded authority makes context responsible for allocating existing prototype evidence rather than directly drawing the output mask.

3.7 Complexity

The architecture uses standard and depthwise convolutions for full spatial representations. KAN operates only in compressed relation spaces (typically 32 channels), and the neck is a single 1×1 projection. The final model contains 5.79M parameters. We report parameter counts and multiply–accumulate operations (MACs) measured from the implemented models under one common 256×256 input. We deliberately label raw profiler counts as MACs rather than mixing them with paper-specific FLOP conventions.

Table 2: Three-seed segmentation results on the three benchmark datasets. Values are mean ± sample standard deviation. Higher is better for IoU and F1; lower is better for HD95. The best result is bold and the second best is underlined. All methods use the same data partitions, checkpoint rule, and metric implementation.

	BUSI			CVC-ClinicDB			GlaS		
Method	IoU ↑	F1 ↑	HD95 ↓	IoU ↑	F1 ↑	HD95 ↓	IoU ↑	F1 ↑	HD95 ↓
U-Net	70.06±0.08	81.58±0.18	9.87±0.27	84.13±1.31	90.99±0.85	2.45±0.67	82.56±0.74	90.33±0.45	1.10±0.00
Attention U-Net	68.60±2.24	80.44±1.63	9.08±0.94	84.72±1.80	91.44±1.08	1.70±1.39	**83.33±1.34**	**90.76±0.85**	1.10±0.10
UNet++	68.55±0.78	80.54±0.56	10.32±0.18	84.51±1.16	91.35±0.75	**1.56±0.44**	82.25±1.31	90.16±0.82	1.06±0.05
UNeXt	67.40±1.32	79.86±0.86	9.39±0.62	77.13±1.33	86.74±0.88	7.28±5.93	76.73±0.82	86.69±0.51	1.29±0.02
U-KAN	72.94±0.69	83.59±0.50	8.38±0.66	83.02±1.70	90.30±1.22	2.46±1.47	82.60±0.96	90.33±0.63	1.10±0.10
Ours	**73.97±1.46**	**84.46±1.07**	**8.01±0.23**	**85.06±0.36**	**91.60±0.23**	2.48±1.58	82.78±1.19	90.49±0.70	**1.01±0.02**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate breast ultrasound, endoscopic polyp, and histopathological gland segmentation on BUSI (Al-Dhabyani et al., 2020), CVC-ClinicDB (Bernal et al., 2015), and GlaS (Sirinukunwattana et al., 2017). BUSI and CVC use fixed 7:2:1 train/validation/test partitions; GlaS uses its official 85/80 split, with 20% of the training portion reserved for validation. We report image-wise IoU, Dice/F1, and HD95 (Karimi & Salcudean, 2020) over seeds $\{42, 2026, 3407\}$ as mean ± sample standard deviation. For each macro-average, we first average the three dataset scores within each seed and then summarize those per-seed averages. All methods share the same partitions, data transformations, 400-epoch budget, validation-based checkpoint rule, and metric implementation; each dataset-specific objective is held fixed across models. Each run uses AdamW on one RTX 4090. The base and reference KAN learning rates are 10^{-4} and 10^{-2}; cosine annealing uses $T_0 = 400$, so no learning-rate restart affects training. Dataset-specific losses, resolutions, augmentation, and optimizer details, including focal Tversky (Abraham & Khan, 2019), are given in Appendix C. The evaluation splits provide neither gradients nor checkpoint-selection signals. Because aggregate benchmark evaluation informed model development, we interpret the reported results as fixed-benchmark performance rather than untouched external-cohort estimates.

Evaluation questions. The experiments ask three questions. First, does the complete architecture maintain competitive behavior across three visually distinct domains rather than relying on one benchmark? Second, how do relation mapping, template conditioning, full-resolution detail, and prototype allocation alter that cross-domain behavior? Third, what parameter, operation, and seed-variation costs accompany the observed accuracy? We answer these questions through a common baseline comparison, retrained component ablations, complexity measurement, and per-seed reporting.

4.2 COMPARISON WITH REPRESENTATIVE METHODS

The comparison is intentionally scoped to representative convolutional, attention-gated, token-MLP, nested, and KAN-based U-shaped models trained from scratch under our common protocol. It is not an exhaustive comparison with foundation-model adaptation or every recent medical segmentation system; accordingly, "best" below always refers to the evaluated implementations.

Table 2 shows the highest BUSI and CVC overlap among the evaluated methods. On BUSI, our IoU exceeds U-KAN by 1.03 points and U-Net by 3.91 points; on CVC, it exceeds Attention U-Net by 0.34 points with a smaller three-seed standard deviation. Attention U-Net leads GlaS IoU/F1 by 0.55/0.27 points, whereas ours has the lowest HD95. Our CVC HD95 is also weaker than the convolutional baselines despite higher overlap, so the gains are complementary rather than uniform. The resulting macro IoU improves over U-KAN by 1.08 points (Table 4). This cross-domain balance is consistent with, but not causal proof of, the intended division between object and boundary evidence.

Table 3: Ablation under the same three-seed protocol. Left: component ablation in mean IoU (%); "MLP both" replaces both KAN relations with approximately parameter-matched MLPs. Right: dataset-agnostic role-fusion controls (mean IoU ± one SD). The ordering study is in Appendix E.

ID	Configuration	BUSI	CVC	GlaS	Avg.
A1	MLP at both relation stages	70.78±3.30	85.45±0.37	80.47±0.33	78.90±1.20
A2	KAN relation at $H/16$ only	70.33±1.99	84.11±2.24	81.19±0.30	78.54±0.30
A3	KAN relation at $H/4$ only	71.73±2.43	82.84±2.42	81.60±0.43	78.73±1.05
A4	Without template conditioning	69.40±2.44	85.10±0.80	80.61±1.77	78.37±1.66
A6	Without full-resolution skip	69.50±2.71	85.27±1.49	81.82±1.69	78.86±1.04
A7	Single prototype pair	69.47±3.16	**86.49±0.35**	81.80±0.39	79.25±0.95
A0	Full model	**73.97±1.46**	85.06±0.36	**82.78±1.19**	**80.60±0.78**

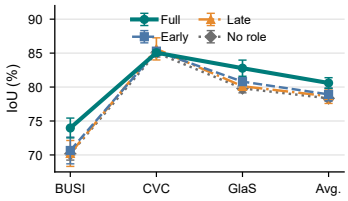

Table 4: Macro-average accuracy and implementation-level complexity. Parameters and raw THOP MACs are measured under the same $1 \times 3 \times 256 \times 256$ input. Custom operators can be incompletely represented by generic profilers, so MACs are not mixed with FLOP numbers quoted from the original papers.

Method	IoU ↑	F1 ↑	HD95 ↓	Params (M) ↓	MACs (G) ↓
U-Net	78.92±0.70	87.63±0.49	4.47±0.32	34.53	65.52
Attention U-Net	78.88±1.49	87.55±1.01	3.96±0.66	34.88	66.63
UNet++	78.43±0.85	87.35±0.55	4.31±0.11	9.16	34.90
UNeXt	73.75±0.61	84.43±0.39	5.99±1.84	**1.47**	**0.57**
U-KAN	79.52±0.72	88.07±0.51	3.98±0.36	6.35	1.75
Ours	**80.60±0.78**	**88.86±0.53**	**3.83±0.60**	5.79	7.19

4.3 Ablation Studies

Component ablations reveal cross-dataset trade-offs. Replacing both relations with MLPs (A1), retaining KAN at one scale (A2–A3), removing template conditioning (A4), or removing the full-resolution skip (A6) all reduce the macro-average. A single prototype pair (A7) improves CVC but loses 4.50 points on BUSI, supporting coordinated evidence allocation rather than universal gains from any isolated module.

The right panel of Table 3 tests the shared role-preservation principle rather than a CVC-specific adjustment. Against early fusion, late fusion, and no role constraint, the full model raises macro IoU by 1.68, 1.95, and 2.26 points. The corresponding macro Dice scores are 88.86, 87.72, 87.51, and 87.31, while macro HD95 is 3.83, 4.03, 4.49, and 3.94 pixels, respectively; the advantage is therefore not confined to IoU. The per-dataset response is deliberately reported rather than hidden by the average: collapsing the roles costs 3.31–3.82 IoU points on BUSI and 1.98–2.97 points on GlaS, whereas CVC remains comparable across variants. This pattern is consistent with preserving local boundary evidence and regional continuity until readout, especially for weak lesion boundaries and fragmented gland topology. It supports a common cross-domain allocation principle, but does not establish a dataset-specific causal mechanism.

4.4 Parameter Count and Stability Analysis

Our 5.79M-parameter model is 8.8% smaller than the evaluated U-KAN and 83.2% smaller than U-Net. Its 7.19G MACs remain above UNeXt and U-KAN, so we claim parameter compactness rather than fastest inference. Seed variation is lowest on CVC (0.36 IoU) and higher on BUSI (1.46) and GlaS (1.19); individual runs appear in Appendix D.

4.5 Qualitative Analysis

Figures 5 and 6 provide complementary qualitative evidence. On BUSI, our model preserves the compact low-contrast lesion without the severe background expansion observed for UNeXt. On CVC, it recovers the elongated target missed by UNeXt and U-KAN and avoids the fragmented response of UNet++. On GlaS, it better maintains gland separation and suppresses spurious internal

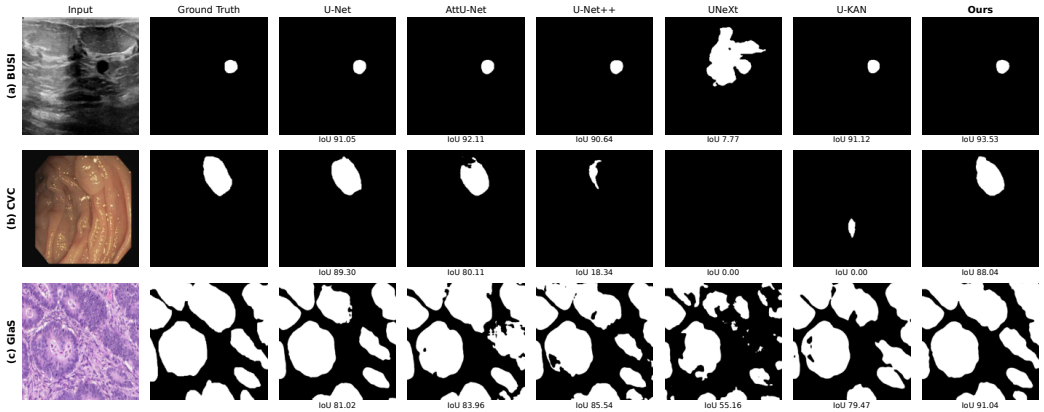

Figure 5: Illustrative comparison on BUSI, CVC, and GlaS. Quantitatively screened examples expose difficult boundaries, false-positive structures, and gland topology, but are not used to compute the reported means.

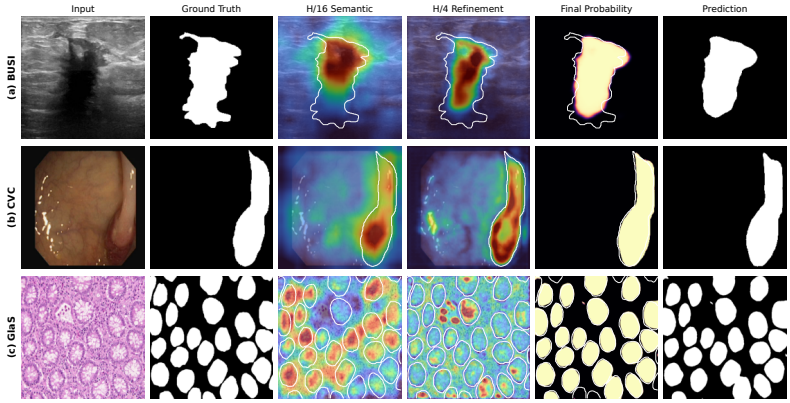

Figure 6: E3-A features at $H/16$ and $H/4$ and the final probability map, shown as qualitative evidence of stage-wise localization rather than an additional selection criterion.

fragmentation. The heatmaps further show a consistent progression: $H/16$ provides broad object-level localization, $H/4$ concentrates responses around regional structure and boundaries, and the final readout suppresses residual off-target activations while forming coherent masks. These selected examples illustrate, rather than prove, the intended stage-wise division of evidence and were not used for model selection.

5 Conclusion

We presented a 5.79M-parameter framework that preserves distinct multiscale evidence roles during relation, transport, and allocation. Under a common three-seed protocol it obtains the highest macro-average IoU and F1 among the evaluated methods. Component and matched fusion controls support this dataset-agnostic design without implying universal per-dataset superiority. Three public 2D benchmarks cannot establish clinical generalization or hardware optimality; broader external evaluation remains future work.

AI Use Statement

Generative AI supported brainstorming, code drafting and debugging, language editing, and LaTeX preparation. The authors reviewed all generated material, conducted every experiment, verified the reported data and references, and take responsibility for the final paper.

Ethics Statement

We use public, de-identified datasets under their respective access terms and collect no new patient data. The model is a research method, not a clinically validated diagnostic system.

Reproducibility Statement

We support reproducibility by specifying the architecture and information flow in Section 3, and the datasets, metrics, and common evaluation protocol in Section 4. Appendix C reports the data splits, preprocessing, optimization settings, learning-rate schedule, checkpoint-selection criterion, and complete training configuration; Appendix D reports every seed-level result. The supplementary material includes the anonymized implementation, environment requirements, training and evaluation scripts, model configurations, representative checkpoints, and instructions for reproducing the main experiments and ablations.

References

Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 683–687, 2019. doi: 10.1109/ISBI.2019.8759329.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. doi: 10.1016/j.dib.2019.104863.

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation versus saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015. doi: 10.1016/j.compmedimag.2015.02.007.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision Workshops*, pp. 205–218, 2022. doi: 10.1007/978-3-031-25066-8_9.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. doi: 10.1016/j.media.2024.103280.

Zehua Cheng, Di Yuan, Wenhu Zhang, and Thomas Lukasiewicz. Effective and efficient medical image segmentation with hierarchical context interaction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9378–9387, 2025.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574–584, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Davood Karimi and Septimiu E. Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 39 (2):499–513, 2020. doi: 10.1109/TMI.2019.2930068.

Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. *Medical Image Analysis*, 67:101851, 2021. doi: 10.1016/j.media.2020.101851.

Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4652–4660, 2025. doi: 10.1609/aaai.v39i5.32491.

Anran Liu, Xiangsheng Huang, Tong Li, and Pengcheng Ma. Co-Net: A collaborative region–contour-driven network for fine-to-finer medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1046–1055, 2022.

Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to upsample by learning to sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6027–6037, 2023.

Yutong Liu, Haijiang Zhu, Mengting Liu, Huaiyuan Yu, Zihan Chen, and Jie Gao. Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3819–3827, 2024. doi: 10.1609/aaai.v38i4.28173.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/afaed89642ea100935e39d39a4da602c-Abstract-Conference.html.

Hao Lu, Wenze Liu, Hongtao Fu, and Zhiguo Cao. FADE: Fusing the assets of decoder and encoder for task-agnostic upsampling. In *European Conference on Computer Vision (ECCV)*, pp. 214–230, 2022a. doi: 10.1007/978-3-031-19812-0_14.

Hao Lu, Wenze Liu, Zixuan Ye, Hongtao Fu, Yuliang Liu, and Zhiguo Cao. SAPA: Similarity-aware point affiliation for feature upsampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022b.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning (MIDL)*, 2018. URL https://arxiv.org/abs/1804.03999.

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. 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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jäger, and Klaus H. Maier-Hein. MedNeXt: Transformer-driven scaling of convnets for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023. doi: 10.1007/978-3-031-43901-8_39.

Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. EGE-UNet: An efficient group enhanced UNet for skin lesion segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023. doi: 10.1007/978-3-031-43901-8_46.

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, 2017. doi: 10.1016/j.media.2016.08.008.

Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5229–5238, 2019.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20730–20740, 2022.

Jeya Maria Jose Valanarasu and Vishal M. Patel. Unext: Mlp-based rapid medical image segmentation network. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 23–33. Springer, 2022. doi: 10.1007/978-3-031-16443-9_3.

Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. CARAFE: Content-aware reassembly of features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3007–3016, 2019.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 325–341, 2018.

Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 173–190, 2020. doi: 10.1007/978-3-030-58539-6_11.

Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020. doi: 10.1109/TMI.2019.2959609.

A Detailed Decoder Architectures

Figures 7–8 expose the complete operators and connections omitted from the main paper for readability. They correspond to the same model used in all experiments and introduce no auxiliary inference branch.

A.1 $H/8$ Regional-Continuity Implementation

We project the upsampled $\mathbf{X}_{16}$ and skip $\mathbf{S}_8$ into a semantic body $\mathbf{B}_8$ and spatial candidate $\mathbf{D}_8$. Their agreement and conflict form

$$\mathbf{R}_8 = \phi_8([\mathbf{B}_8, \mathbf{D}_8, \mathbf{B}_8 \odot \mathbf{D}_8, |\mathbf{B}_8 - \mathbf{D}_8|]). \tag{18}$$

Depthwise branches with dilation $1, 2, 3$ provide local-to-regional context; their fused response conditions a residual refinement that reconstructs $\mathbf{X}_8$. This stage transports regional continuity but does not replace the semantic carrier with the skip candidate.

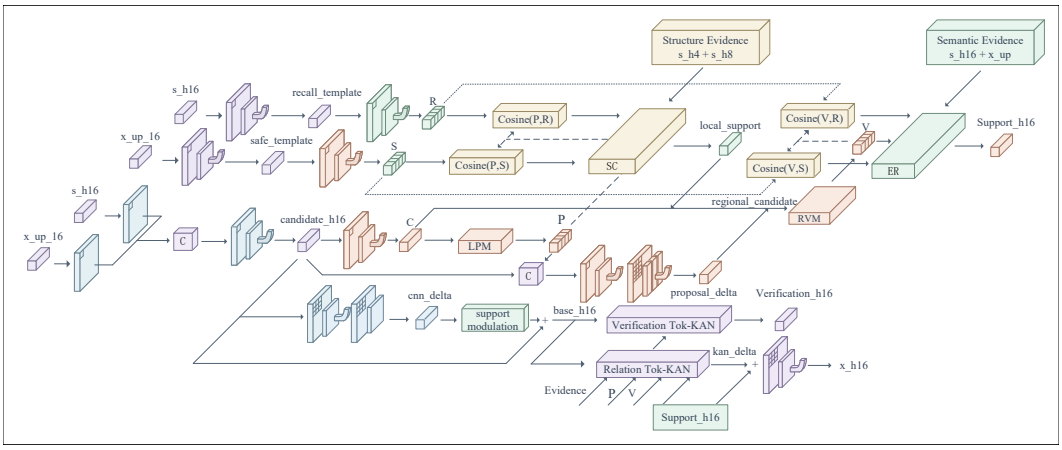

Figure 7: Expanded $H/16$ decoder stage. The upper paths construct cross-scale structural and semantic evidence; the lower paths preserve the convolutional carrier and apply proposal-conditioned nonlinear relation updates.

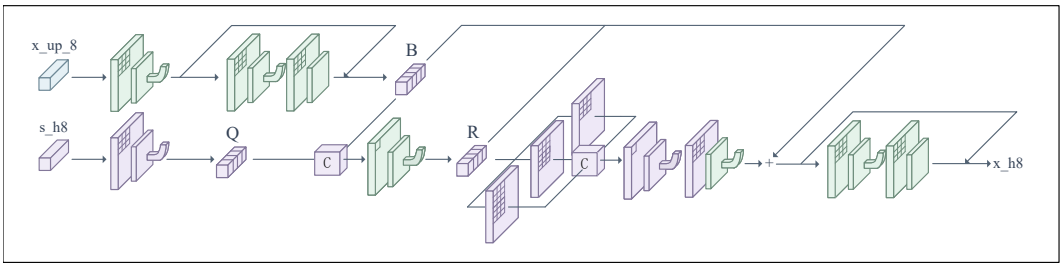

Figure 8: Complete $H/8$ regional-continuity stage. Semantic body and skip candidates interact through agreement, conflict, and local-to-regional context before residual reconstruction.

B Exact Bounded Parameterizations

For completeness, the implementation instantiates the bounded operators in the main text as

$$\mathcal{A}_p(p_l) = \alpha_p(0.25 + 0.75p_l), \tag{19}$$

$$\mathcal{B}_+(a, q, p_l) = \alpha_+ a(1 - p_l)(0.5 + 0.5q), \tag{20}$$

$$\mathcal{B}_-(c, q, p_l) = \alpha_- c[p_l - 0.45]_+(1 - 0.5q), \tag{21}$$

$$b_\gamma(u) = \sigma(u) - 0.5, \qquad b_\beta(u) = 0.25\tanh(u), \tag{22}$$

$$b_\delta(u) = 2\tanh(u). \tag{23}$$

The first operator preserves a nonzero proposal path while increasing its authority with local support; the second and third limit agreement-based reinforcement and conflict-based suppression; the remaining functions cap feature-affine and log-mixture corrections. These numerical bounds are shared by all datasets and are implementation choices rather than theoretically optimal constants. Their purpose is to restrict how strongly a relation path can overwrite its carrier, not to introduce dataset-specific rules.

C Dataset and Training Details

The following details make the reported protocol explicit. The partition seed for BUSI and CVC is 42; for GlaS, the official 85-image training portion is split internally with validation-split seed 2026.

Images are converted to RGB and resized by bilinear interpolation; masks are resized by nearest-neighbor interpolation and binarized at 127. Inputs are normalized with ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. During training, horizontal and

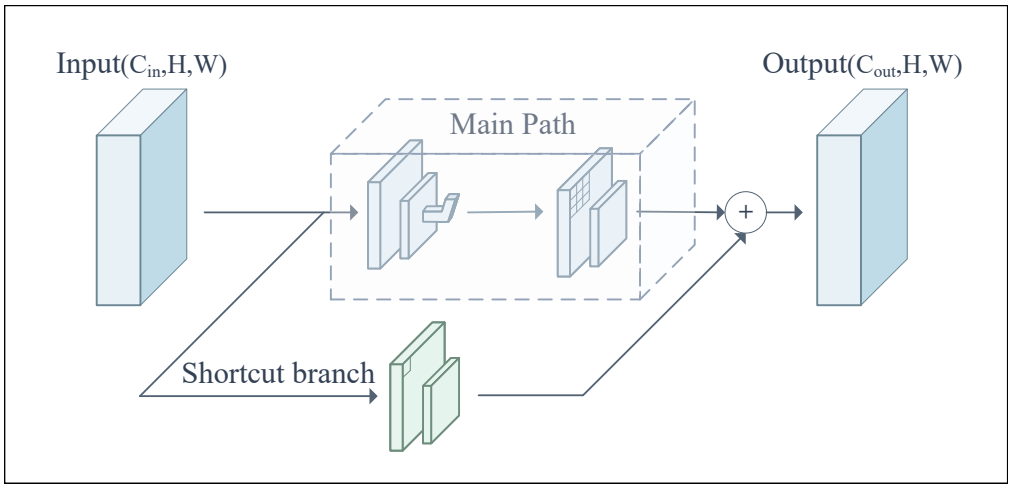

Figure 9: Residual convolutional unit used throughout the hierarchical encoder.

Table 5: Dataset-specific training protocol.

Dataset	Input	Split	Objective
BUSI	256^2	7:2:1	BCE + Dice $(1.0, 1.0)$
CVC	256^2	7:2:1	Focal Tversky $(0.4, 0.6, 1.0)$
GlaS	512^2	$85/80$ + internal 20% val.	Focal Tversky $(0.6, 0.4, 1.0)$

vertical flips are applied independently with probability 0.5, and with probability 0.5 the sample is rotated by one of 90° clockwise, 90° counterclockwise, or 180°, selected uniformly. The same augmentation policy is used for all datasets.

All runs use AdamW, batch size 8, weight decay 10^{-3}, base learning rate 10^{-4}, and a reference KAN learning rate of 10^{-2}. The latter is multiplied by each KAN module's fixed role-specific factor (0.2 or 1.0). Every model is trained for exactly 400 epochs. The cosine scheduler uses $T_0 = 400$, $T_{\text{mult}} = 1$, $\eta_{\min} = 10^{-5}$, and restart-decay factor 0.8. The learning rate therefore follows one cosine cycle; the configured decay is never applied to a subsequent optimization step and no learning-rate reset is used for training. Python, NumPy, PyTorch, CUDA, and data-loader workers are seeded for each run, and deterministic cuDNN execution is enabled. For every seed, the best checkpoint is selected using the validation split; the evaluation split is not used during training or checkpoint selection. The scope of the fixed-benchmark interpretation is stated in Section 4.1.

D Per-Seed Results

E Proposal–Verification Ordering

The primary implementation writes the local proposal before computing the regional verification. To determine whether the empirical gain comes from this exact order, A5 observes the two regions in parallel while retaining the same evidence types and downstream relation mapping. Table 7 shows a dataset-dependent exchange: the parallel variant improves CVC and GlaS but reduces BUSI by 1.86 IoU points and trails the full model by 0.12 points in macro-average IoU. This does not establish sequential processing as universally superior. Rather, it indicates that preserving complementary proposal and verification evidence matters more than enforcing one order.

Table 6: Individual runs of the final model. Reporting all seeds exposes the variation summarized by the main-paper mean ± standard deviation.

	BUSI			CVC-ClinicDB			GlaS		
Seed	IoU	F1	HD95	IoU	F1	HD95	IoU	F1	HD95
42	75.22	85.48	8.2825	84.92	91.54	4.2917	83.91	91.16	1.0000
2026	72.37	83.35	7.8853	85.46	91.86	1.4298	81.54	89.77	1.0414
3407	74.30	84.56	7.8751	84.79	91.42	1.7056	82.87	90.56	1.0000

Table 7: Sensitivity to proposal–verification ordering (IoU, %). Both variants retain the same downstream evidence-relation mechanism.

Observation scheme	BUSI	CVC	GlaS	Macro avg.
Parallel (A5)	72.11±1.01	**86.21±0.52**	**83.11±0.65**	80.48±0.51
Sequential (A0)	**73.97±1.46**	85.06±0.36	82.78±1.19	**80.60±0.78**