

# D<sup>2</sup>-Former: Mixture-Of-Experts Guided Dual Transformer for Multi-Scale Medical Image Segmentation

Md Sohag Mia<sup>\*1,2</sup>

Aya Taourirte<sup>\*1</sup>

Muhammad Abdullah Adnan<sup>2</sup>

Wenlong Ming<sup>†1</sup>

SHUVO2018@NUIST.EDU.CN

202551620018@NUIST.EDU.CN

ADNAN@CSE.BUET.AC.BD

WMING@NUIST.EDU.CN

<sup>1</sup> School of AI, Nanjing University of Information Science and Technology, Nanjing, China

<sup>2</sup> School of Computer Science, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

**Editors:** Accepted for publication at MIDL 2026

## Abstract

Precise delineation of anatomical structures from medical images is critical for clinical diagnosis and treatment planning, yet remains profoundly challenging due to ambiguous boundaries, extreme scale variations, and the heterogeneous appearances of pathological tissues. Current segmentation methods frequently fall short in effectively balancing global contextual understanding with adaptive, multi-scale feature fusion, limiting their robustness across diverse clinical scenarios. To address these limitations, we propose D<sup>2</sup>-Former, a novel encoder-decoder framework that integrates a dual-encoder architecture—combining a Swin Transformer for hierarchical local-global modeling and a DINOv3 foundation model for high-fidelity dense feature extraction—with a Softer Mixture-of-Experts (Softer-MoE) module for input-adaptive feature refinement. Our design further introduces a Spatial-Frequency Gated Channel Attention (SF-GCA) module to fuse complementary encoder representations and a Residual Attention Decoder (RAD) with deep supervision for progressive map reconstruction. Extensive experiments across nine public benchmarks—spanning polyp segmentation, retinal vessel delineation, multi-organ abdominal CT segmentation, and nuclei instance segmentation—demonstrate that D<sup>2</sup>-Former achieves state-of-the-art or highly competitive performance. The model exhibits strong generalization across varied anatomical scales, imaging modalities, and clinical scenarios, underscoring its potential for reliable computer-assisted diagnosis.

**Keywords:** DINOv3, Medical Image Segmentation, Mixture-of-Experts

## 1. Introduction

Medical image segmentation is crucial for computer-assisted diagnosis and treatment yet remains challenging due to low contrast, ambiguous boundaries, and anatomical variations (Rizhi et al., 2025; Ou et al., 2022; Zan et al., 2023; Hatamizadeh et al., 2022; Sharif et al., 2022). Convolutional networks like U-Net (Ronneberger et al., 2015) capture local features but miss long-range dependencies (Hatamizadeh et al., 2022), while transformers like Swin-UNet (Cao et al., 2022b) model global context but lack spatial biases for fine details. Hybrid models with static fusion struggle with medical images’ heterogeneous scales and textures.

---

\* Contributed equally

† Corresponding author

Mixture-of-Experts (MoE) architectures dynamically route information to specialized experts, enabling adaptive capacity without proportional computation (Lepikhin et al., 2020; Riquelme et al., 2021). Softer-MoE (Puigcerver et al., 2024) further allows soft expert combination, producing representations suited for ambiguous boundaries. Meanwhile, self-supervised foundation models like DINOv3 offer high-fidelity features for segmentation (Gao et al., 2025; Yang et al., 2025), but their frozen backbones or simple fusion limit adaptation to medical domains. Related to our goal of complementary encoders, DSU-Net combines DINOv2-guided feature collaboration with a SAM2 backbone using lightweight adapters and cross-layer fusion while largely freezing the foundation backbones (Xu et al., 2025). SAM2-UNeXT (Xiong et al., 2025b) further extends SAM2-UNet (Xiong et al., 2026) by integrating an auxiliary DINOv2 (Oquab et al., 2023) encoder with a dual-resolution strategy and a dense glue layer for encoder fusion. In contrast, D<sup>2</sup>-Former does not adapt a SAM/SAM2 promptable segmentation pipeline; instead, we couple a hierarchical Swin encoder with a DINOv3 encoder and introduce (i) Softer-MoE inside Swin for input-adaptive refinement and (ii) SF-GCA for stage-wise gated fusion tailored to multi-scale medical segmentation.

To address these gaps, we propose D<sup>2</sup>-Former, a Mixture-of-Experts guided dual Transformer framework that couples a Swin Transformer encoder with a DINOv3 foundation-model encoder. A Spatial-Frequency Gated Channel Attention (SF-GCA) module aligns and fuses DINOv3’s high-fidelity dense features with Swin’s hierarchical representations, while Softer-MoE enables adaptive feature refinement within the Swin branch. Unlike prior DINO-based segmentation or MoE vision transformers, D<sup>2</sup>-Former couples a Swin encoder and a DINOv3 foundation encoder via SF-GCA and Softer-MoE specifically for medical image segmentation. The contributions of this work are threefold. First, we introduce a dual-encoder architecture that combines Swin and DINOv3, providing complementary hierarchical local–global modeling and rich semantic feature extraction suitable for diverse medical imaging modalities. Second, we embed Softer-MoE into Swin Transformer blocks, replacing static feed-forward layers with input-dependent expert mixing to better handle large anatomical scale variations and ambiguous boundaries. Third, we design SF-GCA and a Residual Attention Decoder (RAD) with Squeeze and Channel Excitation (CSE), Squeeze and Spatial Excitation (SSE) (Fitzgerald et al., 2024) and deep supervision to adaptively fuse dual-encoder features and progressively recover fine-grained segmentation maps. Extensive experiments across polyp, retinal vessel, multi-organ abdominal CT, and nuclei segmentation benchmarks demonstrate that D<sup>2</sup>-Former achieves improved generalization and consistently enhanced segmentation quality under challenging appearance variations.

## 2. Related Work

Transformers have significantly advanced medical image segmentation by modeling long-range relationships. Early approaches such as TransUNet (Chen et al., 2021a) combined ViT with U-Net, while UNETR (Hatamizadeh et al., 2022) used a pure Transformer encoder for volumetric prediction. Later architectures such as Swin-UNet (Cao et al., 2022b) and PVT (Wang et al., 2021) introduced hierarchical attention for scalable multi-scale feature extraction. Hybrid and dual-path methods (e.g., MedFuseNet (Chen et al., 2025), DTA-SUnet (Ma et al., 2024), GLFNet (Sun et al., 2024)) mitigate Transformers’ limited spatial

priors via multi-branch fusion, but typically rely on static fusion that may under-adapt to large appearance and scale variations.

Foundation models further improve transferability. Self-supervised DINOv3 (Siméoni et al., 2025) provides high-fidelity dense representations; DINO U-Net (Gao et al., 2025) uses DINOv3 as a frozen encoder with adapters, and SegDINO (Yang et al., 2025) employs a lightweight MLP decoder, both demonstrating DINOv3’s transfer learning superiority (Wang et al., 2025). However, they either lack domain adaptation or fail to handle pathological scale variability. In parallel, promptable “segment anything” models have been customized for medical segmentation: SAMed (Zhang and Liu, 2023) adapts SAM via LoRA-based fine-tuning of the image encoder together with the prompt encoder and mask decoder, while SAMed-2 (Yan et al., 2025) builds on SAM-2 (Ravi et al., 2024) by introducing a temporal adapter and a confidence-driven memory mechanism for robust multi-task medical segmentation. Recent SAM2-UNet (Xiong et al., 2026) shows SAM2’s Hiera backbone can serve as a strong encoder for U-shaped models with parameter-efficient adapters, and SAM3-UNet (Xiong et al., 2025a) extends this idea to SAM3 (Carion et al., 2025) using a simple adapter and lightweight U-Net-style decoder for low-cost downstream adaptation.

Mixture-of-Experts (MoE) improves feature specialization and capacity. Sparse MoE (Lepikhin et al., 2020; Riquelme et al., 2021) routes tokens dynamically for efficient scaling. Vision adaptations such as Patcher (Ou et al., 2022) and M<sup>3</sup>ViT (Liang et al., 2022) show MoE gating enhances representation diversity, and Softer-MoE (Puigcerver et al., 2024) softly combines experts for stable representations suitable for ambiguous boundaries and variable scales. These advances motivate integrating Softer-MoE into our dual-encoder framework with SF-GCA for adaptive multi-scale medical segmentation.

### 3. Methods

Our D<sup>2</sup>-Former architecture, illustrated in Figure 1, integrates a DINOv3 and Swin Transformer Module for multi-scale context and our novel Softer-MoE for dynamic feature specialization, which are subsequently fused via SF-GCA modules before being decoded through a U-Net-style decoder with squeeze-and-excitation attention. Below we describe our network’s details.

#### 3.1. Dual Transformer Module

The Dual Transformer Module integrates a Swin Transformer (Base) and a ViT-S+/16 DINOv3 encoder to provide complementary multi-scale features. DINOv3, a self-supervised foundation model, yields high-quality dense representations through discriminative SSL and Gram-anchored training. DINOv3’s first 6 blocks are frozen during training to preserve generic low-level representations and reduce overfitting on medical datasets; higher blocks remain trainable for domain adaptation.

*Swin branch.* Given an image  $X \in \mathbb{R}^{H \times W \times C}$ , non-overlapping  $s_1$ -sized patches form tokens  $X_1 = \text{PatchEmb}(X, s_1)$ , with  $X_1 \in \mathbb{R}^{\frac{H}{s_1} \times \frac{W}{s_1} \times C_1}$ . These pass through hierarchical Swin stages to produce multi-scale features  $F_1 = \{F_s^i\}_{i=1}^4$  (after each patch merging).

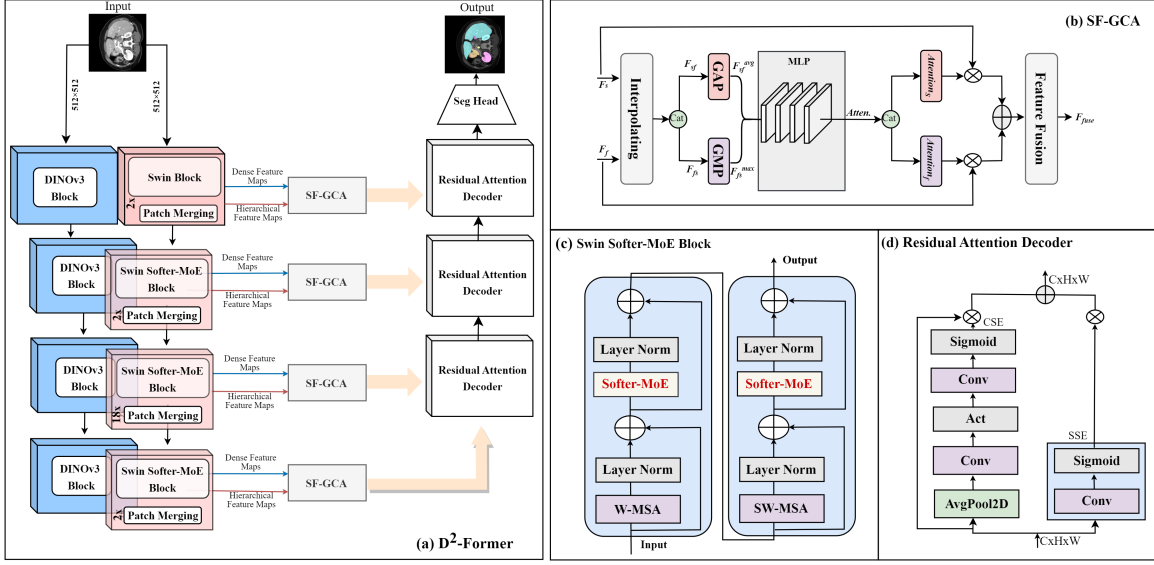


Figure 1: The architecture of our proposed  $D^2$ -Former model. The ‘ $18\times$ ’ indicates that stage-3 has depth 18 in the chosen Swin configuration (depths 2, 2, 18, 2); hence, the corresponding block repeats 18 times. In (b), ‘Cat’ denotes channel-wise concatenation of the aligned features  $[F_s; \hat{F}_f]$ .

*DINOv3 branch.* Following the ViT formulation, the image is patchified with size  $s_2$ , yielding  $Z^{(0)} \in \mathbb{R}^{N \times d}$  where  $N = \frac{H}{s_2} \cdot \frac{W}{s_2}$ . The ViT-S backbone applies  $L$  Transformer blocks,

$$Z^{(\ell)} = B_\ell(Z^{(\ell-1)}), \quad \ell = 1, \dots, L, \quad (1)$$

and we extract token features from four layers  $\{\ell_i\}_{i=1}^4 \subset \{1, \dots, L\}$ , yielding the multi-level DINOv3 representation  $F_2 = \{Z^{(\ell_i)}\}_{i=1}^4$ . In practice we set  $\{\ell_i\} = \{3, 6, 9, 12\}$  to align the four DINOv3 levels with the four Swin stages. Each  $Z^{(\ell_i)}$  is reshaped into a 2D feature map, denoted  $F_f^i$  when used in SF-GCA (Sec. 3.3).

*Fusion and decoding.* At each stage  $i$ , we fuse the Swin feature  $F_s^i$  with the corresponding DINOv3 feature map  $F_f^i$  using SF-GCA, which performs channel projection and spatial alignment (Eq. (7)) followed by gated fusion:

$$F_{\text{out}}^i = \text{SF-GCA}_i(F_s^i, F_f^i), \quad i = 1, \dots, 4. \quad (2)$$

The fused  $\{F_{\text{out}}^i\}_{i=1}^4$  serve as skip features for the residual attention decoder to restore full-resolution segmentation.

*Why dual encoders help?* The dual encoder combines Swin’s strength in modeling hierarchical structures with DINOv3’s robustness in semantic feature extraction. Let  $Y$  be the target mask and  $(F_s^i, F_f^i)$  the stage- $i$  features from Swin and DINOv3. By the chain rule,  $I(Y; F_s^i, F_f^i) = I(Y; F_s^i) + I(Y; F_f^i | F_s^i) \geq I(Y; F_s^i)$ , with strict gain when  $F_f^i$  provides complementary cues. SF-GCA preserves this conditional gain via alignment and channel-wise gating, allowing the decoder to exploit non-redundant multi-scale information.

### 3.2. Swin Softer-MoE Block

We integrate Softer-MoE into Swin Transformer blocks to replace static FFNs with input-adaptive expert mixing. Softer-MoE uses learnable convex (Puigcerver et al., 2024) combinations of experts, enabling content-dependent refinement with minimal computational overhead. We start Softer-MoE from stages 2–4 to avoid the  $O(m^2)$  memory cost at stage 1, because stage 1 has the largest token count, making it memory/compute inefficient and less relevant to semantic scale variation. We emphasize stage 3 (depth 18) since it is the main mid-resolution representation where multi-scale anatomy and boundary ambiguity are most pronounced. We use 8 experts because it provides the best accuracy/compute trade-off in our ablation (Table 7), and soft dense mixing mitigates expert collapse without an extra load-balancing loss.

*Slot Construction and Routing.* Given tokens  $X \in \mathbb{R}^{m \times d}$ , each layer uses  $n$  experts with  $s = m/n$  slots. Slots  $\tilde{X}$  are computed via dispatch weights  $D$

$$D_{ij} = \frac{\exp((X\Phi)_{ij})}{\sum_{i'=1}^m \exp((X\Phi)_{i'j})}, \quad \tilde{X} = D^\top X, \quad (3)$$

where  $\Phi \in \mathbb{R}^{d \times (ns)}$  is learnable. The column-wise softmax encourages each slot to aggregate information from all tokens.

*Expert Processing and Output Combination.* Each slot  $\tilde{X}_i$  is processed by expert  $f_{\lfloor i/s \rfloor}$ . The outputs  $\tilde{Y}$  are merged using combine weights  $C$  and it’s the result of applying softmax over the rows of  $X\Phi$

$$C_{ij} = \frac{\exp((X\Phi)_{ij})}{\sum_{j'=1}^{ns} \exp((X\Phi)_{ij'})}, \quad Y = C\tilde{Y}. \quad (4)$$

*Stage Integration.* Following Swin’s hierarchical reduction  $m/2^{3(i-1)}$ , Softer-MoE is inserted at stages 2–4. We use  $n = 8$  experts and place Softer-MoE layers at (0, 3, 5, 7, 9, 12, 15, 17) in stage 3, and (0, 1) in stages 2 and 4. Stage-3 contains 18 Swin blocks, so we place more Softer-MoE layers there than in stages 2 and 4 (each depth 2). We do not add an explicit load-balancing loss; with 8 experts and dense soft routing, we did not observe expert collapse or training instability.

*Swin Softer-MoE Block.* Each block alternates window-based attention and Softer-MoE:

$$\hat{z}_l = W\text{-MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad z_l = \text{Softer-MoE}(\text{LN}(\hat{z}_l)) + \hat{z}_l, \quad (5)$$

$$\hat{z}_{l+1} = SW\text{-MSA}(\text{LN}(z_l)) + z_l, \quad z_{l+1} = \text{Softer-MoE}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1}. \quad (6)$$

This preserves Swin’s local–global hierarchy while enabling dynamic, token-dependent expert routing. The Swin Softer-MoE block is depicted in Figure 1(c).

### 3.3. Spatial-Frequency Gated Channel Attention

We introduce a Spatial-Frequency Gated Channel Attention (SF-GCA) module to fuse the complementary representations from the dual encoders (Figure 1(b)). At encoder stage  $i$ , the Swin encoder yields a spatial feature map  $F_s^i \in \mathbb{R}^{C_s \times H \times W}$ , while DINOv3 provides a globally contextual frequency-rich feature map  $F_f^i \in \mathbb{R}^{C_f \times H' \times W'}$ . For readability, we omit the stage index  $i$  below.

To match dimensions,  $F_f$  is projected from  $C_f$  to  $C_s$  channels via a  $1 \times 1$  convolution, followed by GroupNorm and spatial alignment:

$$\hat{F}_f = \text{Align}(\text{GN}(\text{Conv}_{1 \times 1}(F_f))). \quad (7)$$

Here  $\text{Align}(\cdot)$  denotes 2D bilinear interpolation to match the spatial resolution of  $F_s$ . The aligned features are concatenated and processed through a dual-branch channel attention mechanism using global average pooling (GAP) and global max pooling (GMP). Both descriptors pass through a shared two-layer MLP (reduction ratio  $r = 16$ ) with sigmoid activation:

$$A = \sigma\left(\text{MLP}(\text{GAP}([F_s; \hat{F}_f])) + \text{MLP}(\text{GMP}([F_s; \hat{F}_f]))\right). \quad (8)$$

Here  $[\cdot; \cdot]$  denotes channel-wise concatenation. Since  $[F_s; \hat{F}_f] \in \mathbb{R}^{2C_s \times H \times W}$ , the attention vector satisfies  $A \in \mathbb{R}^{2C_s}$ . We split it as  $A = [A_s; A_f]$  with  $A_s, A_f \in \mathbb{R}^{C_s}$  (first  $C_s$  and last  $C_s$  ( $C_f$ ) channels), and apply channel-wise gating:  $\tilde{F}_s = A_s \odot F_s$  and  $\tilde{F}_f = A_f \odot \hat{F}_f$ , where  $\odot$  denotes per-channel multiplication. Fusion is then performed through learnable adaptive weighting:

$$F_{fused} = \frac{\alpha \tilde{F}_s + \beta \tilde{F}_f}{\alpha + \beta}, \quad (9)$$

where  $\alpha$  and  $\beta$  are trainable scalars. This enables dynamic balancing between spatial detail and frequency-domain context, improving cross-encoder consistency across scales. We instantiate SF-GCA at all four encoder stages so that fused features are available at each scale for the decoder.

### 3.4. Residual Attention Decoder

We use a Residual Attention Decoder (RAD) (Figure 1(d)) to progressively reconstruct dense segmentation maps. Each stage upsamples the incoming feature map to the skip resolution, concatenates the encoder shortcut (when present), and applies a Spatial and Channel Squeeze and Excitation (SCSE) module to recalibrate channel and spatial responses. Two consecutive  $3 \times 3$  layers follow for local aggregation, after which a second SCSE refines the output. The final segmentation head applies a Residual Block (RB) with Mish activation and GroupNorm for last-stage refinement prior to upsampling and  $1 \times 1$  prediction. The RB is therefore a post-decoder refinement module.

During training, we employ deep supervision with two auxiliary output branches connected to intermediate decoder stages. The auxiliary losses are weighted by factors 0.2 and 0.1 respectively (reduced to 0.1 and 0.05 during warmup), encouraging the network to learn discriminative features at multiple scales while stabilizing gradient flow.

## 4. Experiments

Our D<sup>2</sup>-Former model is evaluated on multiple medical image segmentation datasets, with implementation details and further results provided in subsequent sections.

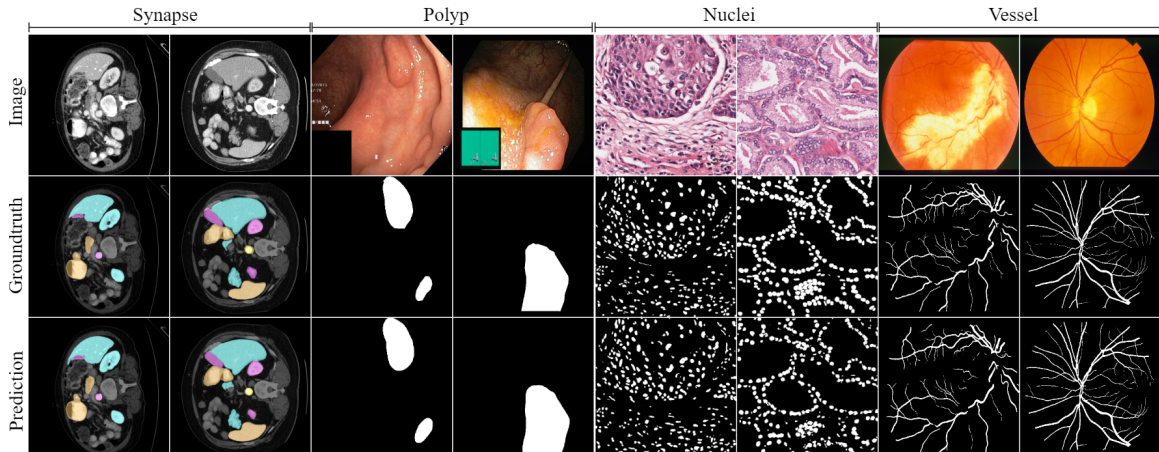


Figure 2: Qualitative results on synapse, polyp, nuclei, and retinal vessel datasets.

Table 1: Comparison to the SOTA methods on four polyp datasets.

Methods	Kvasir-SEG		CVC-ClinicDB		CVC-300		CVC-ColonDB	
	Dice $\uparrow$	mIoU $\uparrow$	Dice $\uparrow$	mIoU $\uparrow$	Dice $\uparrow$	mIoU $\uparrow$	Dice $\uparrow$	mIoU $\uparrow$
U-Net (Ronneberger et al., 2015)	81.8	74.6	82.3	75.5	71.0	62.7	51.2	44.4
PraNet (Fan et al., 2020)	89.8	84.0	89.9	84.9	87.1	79.7	70.9	64.0
Polyp-PVT (Dong et al., 2023)	91.7	86.4	93.7	88.9	90.0	83.3	80.8	72.7
SegDINOv3 (Yang et al., 2025)	87.6	80.6	94.0	88.2	90.2	85.0	81.2	73.2
PAAN (Yi et al., 2024)	<b>94.2</b>	<b>89.7</b>	93.4	88.4	<b>92.6</b>	<b>86.9</b>	78.6	71.6
VM-UNet (J and S, 2024)	91.3	85.6	92.6	87.1	88.6	81.8	79.8	71.2
QueryNet (Chai et al., 2024)	93.2	<u>88.3</u>	<u>94.2</u>	<b>89.4</b>	92.0	85.9	<u>82.7</u>	<u>75.9</u>
SSformer (Wang et al., 2022)	92.6	87.4	92.7	87.6	89.5	82.7	80.2	72.1
ColnNet (Jain et al., 2023)	92.6	87.2	93.0	88.7	90.9	86.3	79.7	72.9
<b>D<sup>2</sup>-Former</b>	<u>93.4</u>	87.3	<b>94.2</b>	<u>88.9</u>	<u>92.0</u>	<u>86.1</u>	<b>82.8</b>	<b>76.3</b>

#### 4.1. Experimental Setup, Datasets, and Metrics

Our model was implemented in PyTorch and trained on an NVIDIA RTX 4090 GPU using  $512 \times 512$  input images with a batch size of 8 for 100 epochs. We employed flipping augmentation and used the AdamW optimizer with an initial learning rate of  $1e-4$ . Evaluation for polyp segmentation was performed on the Kvasir-SEG (Jha et al., 2020), CVC-ClinicDB (Bernal et al., 2015), and CVC-ColonDB (Tajbakhsh et al., 2015) datasets. We further conducted extensive experiments on several additional public benchmarks: the DRIVE (Staal et al., 2004) and STARE (Hoover et al., 2000) datasets for retinal vessel segmentation, the Synapse dataset (Chen et al., 2021b) for multi-organ abdominal CT segmentation (Synapse is handled 2D slice-wise with the standard split), and the MoNuSeg (Kumar et al., 2020) and CryoNuSeg (Mahbod et al., 2021) datasets for nuclei instance segmentation in histopathology. Performance was assessed using the Dice, mIoU, S-measure ( $S_\alpha$ ), F-measure ( $F_\beta^\omega$ ), E-measure ( $E_p^{\max}$ ), F1-Score, Acc, MAE, AJI, PQ, SP, and SE metrics. More details are in Appendix A.

Table 2: Quantitative results comparison on multi-organ abdominal CT dataset. Kidney Left (KL), Kidney Right (KR), Gallbladder (GB).

Methods	Dice $\uparrow$	mIoU $\uparrow$	Aorta	GB	KL	KR	Liver	Pancreas	Spleen	Stomach
AttnUNet	71.70	61.38	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
SSFormer	78.01	67.23	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
PolypPVT	78.08	67.43	82.34	66.14	81.21	73.78	94.37	59.34	88.05	79.40
TransUNet	77.61	67.32	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75.00
SwinUNet	77.58	66.88	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
PVT-CASCADE	81.06	70.88	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
TransCASCADE	82.68	73.48	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52
PVT-EMCAD-B2	83.63	74.65	88.14	68.87	88.08	84.10	95.26	68.51	92.17	83.92
<b>D<sup>2</sup>-Former</b>	<b>85.53</b>	<b>77.32</b>	<b>91.23</b>	<b>71.94</b>	<b>89.90</b>	<b>86.21</b>	<b>96.11</b>	<b>70.03</b>	<b>94.10</b>	<b>84.74</b>

Table 3: Comparison to the SOTA methods on nuclei segmentation datasets.

Methods	MoNuSeg			CryoNuSeg		
	Dice $\uparrow$	AJI $\uparrow$	PQ $\uparrow$	Dice $\uparrow$	AJI $\uparrow$	PQ $\uparrow$
SAMUS (Lin et al., 2024)	83.26	67.98	63.93	84.40	51.25	48.35
SMMILe (Lu et al., 2023)	<b>89.20</b>	n/a	n/a	<u>87.60</u>	n/a	n/a
DenseU-Net (Kiran et al., 2022)	83.21	<b>78.61</b>	n/a	83.50	<b>79.42</b>	n/a
NucleiSeg (Swain et al., 2024)	<u>89.02</u>	n/a	n/a	n/a	n/a	n/a
NuSegDG (Lou et al., 2024)	86.37	69.81	<u>68.88</u>	84.59	53.33	<u>49.11</u>
<b>D<sup>2</sup>-Former</b>	88.93	<u>71.02</u>	<b>70.13</b>	<b>87.72</b>	<u>70.10</u>	<b>63.21</b>

Table 4: Quantitative results on retinal vessel segmentation datasets.

Methods	DRIVE				STARE			
	F1 Score $\uparrow$	Acc $\uparrow$	SP $\uparrow$	SE $\uparrow$	F1 Score $\uparrow$	Acc $\uparrow$	SP $\uparrow$	SE $\uparrow$
TCDD-UNet (Nianzhu et al., 2024)	<u>82.65</u>	<u>96.98</u>	<u>98.12</u>	<u>84.21</u>	<u>81.63</u>	97.40	98.37	<b>82.24</b>
U-Net++ (Zhou et al., 2018)	81.92	96.88	97.43	82.51	78.59	<u>97.57</u>	98.51	76.83
DUNet (Qiangguo et al., 2019)	82.37	95.66	97.92	79.70	81.43	96.41	<u>98.81</u>	73.70
MambaUNet (Wang et al., 2024)	81.92	95.51	97.73	80.48	78.96	96.11	98.75	73.19
AttUKAN (Zeng et al., 2025)	82.50	95.49	97.21	83.93	81.14	96.43	98.68	76.88
<b>D<sup>2</sup>-Former</b>	<b>83.21</b>	<b>97.19</b>	<b>98.31</b>	<b>84.40</b>	<b>82.11</b>	<b>97.84</b>	<b>98.83</b>	<u>79.87</u>

## 4.2. Experimental Results

**Polyp Segmentation:** The proposed D<sup>2</sup>-Former achieves competitive performance across four polyp segmentation benchmarks (Table 9). It obtains the highest Dice/IoU on CVC-ClinicDB (94.2/88.9) and CVC-ColonDB (82.8/76.3). On Kvasir-SEG, it ranks second (93.4/87.3) and remains highly competitive on CVC-300 (92.0/86.1). These results demonstrate D<sup>2</sup>-Former’s strong generalization. A detailed analysis can be found in Appendix B.1.

**Nuclei Segmentation:** D<sup>2</sup>-Former delivers competitive nuclei segmentation performance, as summarized in Table 3. On CryoNuSeg, it achieves the highest Dice of 87.7, along with strong AJI (70.1) and PQ (63.2), outperforming existing methods across all reported metrics. For the MoNuSeg dataset, D<sup>2</sup>-Former attains a Dice of 88.9 with improved AJI (71.0) and PQ (70.1), remaining highly competitive with leading approaches such as SMMILe and

Table 5: Comprehensive reproducibility analysis across 5 independent training runs. Results are reported as mean ( $\mu$ )  $\pm$  standard deviation ( $\sigma$ ) for multiple evaluation metrics.

Dataset	$\mu \pm \sigma$	Dice $\uparrow$	mIoU $\uparrow$	AJI $\uparrow$	PQ $\uparrow$	F1 Score $\uparrow$	Acc $\uparrow$	SP $\uparrow$	SE $\uparrow$
Kvasir	$\mu$	93.34	87.22	n/a	n/a	n/a	n/a	n/a	n/a
	$\sigma$	0.12	0.15	n/a	n/a	n/a	n/a	n/a	n/a
CVC-ClinicDB	$\mu$	94.08	88.81	n/a	n/a	n/a	n/a	n/a	n/a
	$\sigma$	0.19	0.18	n/a	n/a	n/a	n/a	n/a	n/a
Synapse	$\mu$	85.44	77.20	n/a	n/a	n/a	n/a	n/a	n/a
	$\sigma$	0.20	0.23	n/a	n/a	n/a	n/a	n/a	n/a
MoNuSeg	$\mu$	88.84	n/a	70.80	69.98	n/a	n/a	n/a	n/a
	$\sigma$	0.12	n/a	0.31	0.18	n/a	n/a	n/a	n/a
CryoNuSeg	$\mu$	87.59	n/a	70.02	63.11	n/a	n/a	n/a	n/a
	$\sigma$	0.21	n/a	0.13	0.14	n/a	n/a	n/a	n/a
DRIVE	$\mu$	n/a	n/a	n/a	n/a	83.10	97.14	98.19	84.28
	$\sigma$	n/a	n/a	n/a	n/a	0.16	0.10	0.22	0.26
STARE	$\mu$	n/a	n/a	n/a	n/a	82.01	97.75	98.75	79.80
	$\sigma$	n/a	n/a	n/a	n/a	0.21	0.18	0.16	0.10

NucleiSeg. These results demonstrate that D<sup>2</sup>-Former is a robust and effective model for nuclei segmentation across diverse histopathological datasets.

**Retinal Vessel Segmentation:** D<sup>2</sup>-Former demonstrates strong performance on retinal vessel segmentation, enabling accurate delineation of fine vascular structures. As reported in Table 4, our model achieves the highest F1 scores on both DRIVE (83.21) and STARE (82.11), together with superior accuracy, specificity, and sensitivity compared to recent CNN-, Mamba-, and KAN-based baselines. These results highlight the robustness of D<sup>2</sup>-Former across retinal datasets, maintaining reliable performance on thin and low-contrast vessel structures.

**Multi-organ Abdominal CT Segmentation:** Our model exhibits exceptional generalization on the challenging multi-class Synapse dataset, a standard benchmark for abdominal CT segmentation with eight distinct organ classes. As shown in Table 2, our model achieves the highest mean Dice of 85.53 and mIoU of 77.32, significantly outperforming recent strong baselines such as TransCASCADE (Dice 82.68) (Rahman and Marculescu, 2023), PVT-CASCADE (Dice 81.06) (Titoriya and Singh, 2023), SwinUNet (Dice 77.58) (Cao et al., 2022a), and PVT-EMCAD-B2 (Dice 83.63) (Rahman et al., 2024). D<sup>2</sup>-Former demonstrates superior performance across all organs, particularly excelling in anatomically complex structures like the pancreas (70.03 Dice) and gallbladder (71.94 Dice), where other models often struggle. This indicates that the DINOv3-driven dual encoder, combined with adaptive routing via Softer-MoE, effectively captures both high-fidelity semantic features and fine-grained structural details, enabling robust segmentation across highly variable organ appearances and scales. Qualitative segmentation results across four distinct biomedical imaging domains—polyp, nuclei, retinal vessel, and multi-organ CT—are visualized in Figure 2, and additional predicted images from synapse and retinal vessel are shown in Appendix B.

**Statistical Validation and Reproducibility Analysis:** We evaluate reproducibility by repeating each experiment 5 times with different random seeds while keeping data splits and hyperparameters fixed. Table 5 reports the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of all metrics across runs. The low variance observed across datasets (0.12–0.21 for Dice and 0.15–0.23 for mIoU) indicates strong robustness to initialization. In addition, we performed

Table 6: Ablation study on dual-encoder components and fusion strategies. Encoder-Decoder (ED).

Configuration	Kvasir-SEG	CVC-ClinicDB	CVC-ColonDB	DRIVE	STARE	Synapse
Baseline (Swin-ED)	90.85	92.50	80.32	81.21	79.38	80.74
+ DINOv3 (concat)	92.22	93.41	81.90	82.76	81.37	83.81
+ SF-GCA	93.08	93.98	82.36	83.10	82.02	84.67
+ RAD (ours)	<b>93.42</b>	<b>94.21</b>	<b>82.83</b>	<b>83.21</b>	<b>82.11</b>	<b>85.53</b>

Table 7: Ablation on Softer-MoE variants and stage placements.

MoE Configuration	Kvasir	CVC-ClinicDB	Synapse	DRIVE
No MoE (baseline FFN)	93.12	93.79	84.22	83.03
MoE (Hwang et al., 2023)	93.22	93.98	85.18	83.15
Softer-MoE (stages 3–4)	93.17	93.96	84.95	83.08
Softer-MoE (stages 2–4)	93.20	94.00	85.03	83.10
Softer-MoE (4 experts)	93.23	94.06	85.11	83.18
Softer-MoE (8 experts)	93.42	94.21	85.53	83.21

paired  $t$ -tests over the 5 matched-seed runs ( $n = 5$ ) comparing D<sup>2</sup>-Former to the Swin-ED baseline; improvements are significant across all reported metrics ( $p < 0.001$ ).

### 4.3. Ablation Study

To comprehensively evaluate the contribution of each proposed component, we conduct systematic ablation experiments on the Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, DRIVE, STARE, and Synapse datasets. The baseline is a Swin-Transformer encoder-decoder with standard FFN blocks. All models are trained under identical settings (Sec. 4.1), and performance is measured by Dice score (%) and F1-score (for retinal vessel datasets).

#### 4.3.1. IMPACT OF DUAL ENCODER AND FUSION MODULES

We first investigate the contribution of the dual-encoder architecture and the proposed fusion mechanisms. The ablation study in Table 6 demonstrates the progressive improvement brought by each component of our dual-encoder design. The addition of the DINOv3 branch with simple concatenation enhances performance across all datasets, demonstrating dual-encoder superiority and confirming that the foundation model’s high-fidelity dense features effectively complement the Swin Transformer’s hierarchical local-global modeling. Replacing this concatenation with our Spatial-Frequency Gated Channel Attention (SF-GCA) module yields further gains, as it enables adaptive, channel-wise fusion of spatial details and frequency-domain context through learnable gating. Finally, integrating the Residual Attention Decoder (RAD) achieves the best results, validating that our decoder’s multi-stage refinement with skip connections and attention mechanisms is essential for precise boundary delineation and robust generalization across diverse anatomical structures.

#### 4.3.2. ROLE OF SOFTER-MOE IN ADAPTIVE FEATURE REFINEMENT

The ablation results in Table 7 confirm that replacing the static FFN with any MoE variant consistently improves performance across all datasets, validating the benefit of dy-

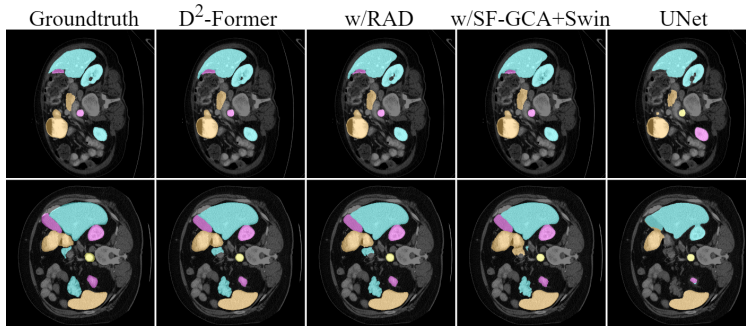


Figure 3: Visual ablation study on the synapse dataset.

Table 8: Ablation on decoder components and supervision.

Decoder Variant	Kvasir	Synapse	CryoNuSeg
Plain U-Net decoder	90.54	82.06	84.82
+ Skip connections only	90.89	82.51	84.99
+ SCSE modules	92.93	84.76	86.73
+ Residual Blocks (RB)	93.13	85.41	87.42
+ Deep supervision (full RAD)	<b>93.42</b>	<b>85.53</b>	<b>87.71</b>

dynamic expert routing for multi-scale anatomical structures. While base MoE and our initial Softer-MoE configurations (2 experts, stages 3-4 or 2-4) provide moderate gains, the optimal configuration employs 8 experts across stages 2-4, which yields the highest Dice scores (e.g., +0.30 on Kvasir and +1.31 on Synapse over baseline). This setup allows the model to apply soft, input-dependent expert blending from mid-to-high feature levels, enhancing adaptive refinement of both local boundaries and global context without over-specializing on low-level patterns.

#### 4.3.3. EFFECT OF DECODER DESIGN AND SUPERVISION STRATEGY

We evaluate the contribution of the Residual Attention Decoder (RAD) and the deep supervision scheme. Table 8 presents the results. While the plain U-Net decoder effectively combines multi-scale features, it struggles to refine complex anatomical boundaries due to limited representational capacity. The introduction of SCSE modules addresses this by performing channel-spatial recalibration, yielding a marked improvement (e.g., +2.87 on Synapse) through better focus on salient structures. The Residual Block (RB) in the segmentation head further sharpens boundary delineation via local residual refinement. Finally, deep supervision stabilizes learning and improves performance, confirming RAD’s effectiveness in segmenting complex anatomical regions (Figure 3).

#### 4.3.4. EFFECT OF DINO BACKBONE VARIANTS

Table 9 reports a controlled comparison between DINOv2 and DINOv3 backbones within the proposed D<sup>2</sup>-Former framework. Across all benchmarks, DINOv3 variants consistently outperform their DINOv2 counterparts, demonstrating superior transferability of DINOv3’s high-fidelity dense representations to medical image segmentation. In particular, DINOv3-

Table 9: Comparison to the different DINO variant models.

Models	Kvasir-SEG		Synapse		DRIVE		MoNuSeg	
	Dice $\uparrow$	mIoU $\uparrow$	Dice $\uparrow$	mIoU $\uparrow$	F1 Score $\uparrow$	Acc $\uparrow$	Dice $\uparrow$	AJI $\uparrow$
DINOv2-S	92.04	86.26	85.02	76.75	82.13	96.31	87.56	69.23
DINOv2-L	92.87	87.03	85.32	77.11	82.97	97.08	88.52	70.50
DINOv3-S+	93.42	87.31	85.53	77.32	83.21	97.19	88.93	71.02
DINOv3-L	93.81	88.79	85.92	78.98	84.11	97.63	89.21	71.15

Table 10: Overall model complexity.

Model	Parameters (M) $\downarrow$	FLOPs (G) $\downarrow$	Inf-GPU (ms) $\downarrow$	Inf-CPU (ms) $\downarrow$	Model Size (MB) $\downarrow$
TransUNet	86	11.7	22	280	328
AttnUNet	31.5	3.9	11	135	120
D <sup>2</sup> -Former	119.9	66.21	84.8	265	457

S+ improves Dice and mIoU on Kvasir-SEG and Synapse, as well as F1-score and AJI on DRIVE and MoNuSeg, respectively. Although DINOv3-L achieves the best absolute performance, the improvement over DINOv3-S+ is marginal, while incurring significantly higher computational cost. Therefore, we adopt DINOv3-S+ for all experiments and ablation studies, as it provides the best balance between segmentation accuracy and efficiency.

#### 4.3.5. COMPUTATIONAL EFFICIENCY ANALYSIS

We compare the computational complexity of our model with representative baselines in Table 10. D<sup>2</sup>-Former has 119.9M parameters and requires 66.21G FLOPs, making it larger and computationally heavier than lightweight models such as AttnUNet and TransUNet, and resulting in higher GPU and CPU inference latency. This overhead mainly stems from the dual-encoder architecture and the integration of Softer-MoE, which enhance representational capacity for complex anatomical structures. Despite the increased complexity, D<sup>2</sup>-Former remains practically viable, achieving efficient GPU inference (84.8 ms per 256 $\times$ 256 image on an NVIDIA RTX 4090), and the added cost consistently translates into superior segmentation accuracy across diverse datasets.

## 5. Conclusion

Accurate medical image segmentation demands robust modeling of ambiguous boundaries and multi-scale anatomical variations. We present D<sup>2</sup>-Former, a dual-encoder framework that integrates a Swin Transformer for hierarchical structural modeling with a DINOv3 foundation model for high-fidelity semantic features. To adaptively refine features, we introduce Softer-MoE blocks that softly route tokens to specialized experts, along with a Spatial-Frequency Gated Channel Attention (SF-GCA) module and a Residual Attention Decoder (RAD) for precise map reconstruction. Evaluated across nine public benchmarks covering polyp, retinal vessel, multi-organ CT, and nuclei segmentation, D<sup>2</sup>-Former achieves state-of-the-art or highly competitive performance, demonstrating strong generalization across imaging modalities and anatomical scales. Our work establishes that Mixture-of-Experts guided dual Transformers offer a scalable and effective paradigm for clinical segmentation tasks. Future work will extend D<sup>2</sup>-Former to 3D medical volumes, exploring volumetric Softer-MoE designs and spatio-temporal attention mechanisms for enhanced cross-modal understanding.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62401272 and 82441029).

## References

- Jorge Bernal, FJavier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dovamaps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, page 99–111, 2015.
- 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 *In European conference on computer vision*, pages 205–218, 2022a.
- Hu Cao, Yueyue Wang, Joy Chen Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *in ECCV*, 2022b.
- Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL <https://arxiv.org/abs/2511.16719>.
- Jiaxing Chai, Zhiming Luo, Jianzhe Gao, Licun Dai, Yingxin Lai, and Shaozi Li. QueryNet: A Unified Framework for Accurate Polyp Segmentation and Detection . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008. Springer Nature Switzerland, 2024.
- Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv*, 2021a.
- Jieneng Chen et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021b.
- Ruiyuan Chen, Saiqi He, Hongsheng Lu, and Zhaohui Yang. Medfusenet: fusing local and global deep feature representations with hybrid attention mechanisms for medical image segmentation. *Sci Rep*, 2025.
- Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research*, 2023.

- Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, and Geng Chen. Pranel: Parallel reverse attention network for polyp segmentation. In *in MICCAI*, 2020.
- Kerr Fitzgerald, Jorge Bernal, Aymeric Histace, and Bogdan J. Matuszewski. Polyp segmentation with the fcb-swinv2 transformer. *IEEE Access*, 12:38927–38943, 2024.
- Yifan Gao, Haoyue Li, Feng Yuan, Xiaosong Wang, and Xin Gao. Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. *ArXiv*, abs/2508.20909, 2025. URL <https://api.semanticscholar.org/CorpusID:280949513>.
- Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *in WACV*, 2022.
- Adam Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210, 2000.
- Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, and Zilong Wang. Tutel: Adaptive mixture-of-experts at scale. In *Proceedings of Machine Learning and Systems 5*, pages 269–287, 2023.
- Ruan J and Xiang. S. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv*, 2024.
- Samir Jain, Rohan Atale, Anubhav Gupta, Utkarsh Mishra, Ayan Seal, Aparajita Ojha, Joanna Jaworek-Korjakowska, and Ondrej Krejcar. Coinnet: A convolution-involution network with a novel statistical attention for automatic polyp segmentation. *IEEE Transactions on Medical Imaging*, 42:3987–4000, 2023. doi: 10.1109/TMI.2023.3320151.
- Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. *in ICMM*, pages 451—462, 2020.
- Iqra Kiran, Basit Raza, Areesha Ijaz, and Muazzam A. Khan. Denseres-unet: Segmentation of overlapped/clustered nuclei from multi organ histopathology images. *Computers in Biology and Medicine*, 143:105267–, 2022.
- Neeraj Kumar et al. A multi-organ nucleus segmentation challenge. In *IEEE Transactions on Medical Imaging*, volume 39, pages 1380–1391, 2020.
- Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *in ICLR*, 2020.
- Hanxue Liang, Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, and Zhangyang Wang. M<sup>3</sup>vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *in NeurIPS*, 2022.

- Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In *proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.
- Zhenye Lou, Qing Xu, Zekun Jiang, Xiangjian He, Zhen Chen, Yi Wang, Chenxin Li, Maggie M He, and Wenting Duan. Nusegdg: Integration of heterogeneous space and gaussian kernel for domain-generalized nuclei segmentation. *arXiv preprint arXiv:2408.11787*, 2024.
- Cheng Lu, Zhe Feng, Zhihua Liu, Jun Cheng, Chu Han, Li Li, Feng Hou, Cheng Liang, Hao Wang, Zhenbing Liu, Xipeng Pan, Xiang Chen, and Rushi Lan. SMILE: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations. *Medical Image Analysis*, 88:102867, 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102867.
- Bo Ma, Qian Sun, Ze Ma, Baosheng Li, Qiang Cao, Yungang Wang, and Gang Yu. Dtasunet: a local and global dual transformer with the attention supervision u-network for brain tumor segmentation. *Sci Rep*, 2024.
- Amirreza Mahbod et al. Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. *Computers in Biology and Medicine*, 132:104349, 2021.
- Lv Nianzhu, Xu Li., and Chen. Teddu-net: combining transformer and convolutional dual-path decoding u-net for retinal vessel segmentation. *Sci Rep 14*, 25978:149–162, 2024.
- Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.
- Yanglan Ou, Ye Yuan, Xiaolei Huang, Stephen T. C. Wong, John Volpi, James Z. Wang, and Kelvin Wong. Patcher: Patch transformers with mixture of experts for precise medical image segmentation. In *in MICCAI*, pages 475–484, 2022.
- Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *ICLR*, 2024.
- Jin Qiangguo, Zhaopeng Meng, Tuan D. Pham, Qi Chen, Leyi Wei, , and Ran Su. Dunet: A deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, 178: 149–162, 2019.
- Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6211–6220, 2023. doi: 10.1109/WACV56688.2023.00616.

- 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*, pages 11769–11779, 2024.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL <https://arxiv.org/abs/2408.00714>.
- Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. in *NeurIPS*, 2021.
- Ding Rizhi, Lu Hui, and Liu M. Densformer-moe: A dense transformer foundation model with mixture of experts for multi-task brain image analysis. *IEEE TIM*, 2025.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *in MICCAI*, 2015.
- Muhammad Hamza Sharif, Dmitry Demidov, Asif Hanif, Mohammad Yaqub, and Min Xu. Transresnet: Integrating the strengths of vits and cnns for high resolution medical image segmentation via feature grafting. In *in BMVC*, 2022.
- Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dino3, 2025. URL <https://arxiv.org/abs/2508.10104>.
- Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.
- Shiyao Sun, Chong Fu, Sen Xu, Yingyou Wen, and Tao Ma. Glnet: Global-local fusion network for the segmentation in ultrasound images. *Comput Biol Med*, 2024.
- Bishal Ranjan Swain, Kyung Joo Cheoi, and Jaepil Ko. Nuclei segmentation in histopathological images with enhanced u-net3+. In *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, volume 250, pages 1513–1530, 2024.
- Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE TMI*, page 630–644, 2015.
- A. K. Titoriya and M. P. Singh. Pvt-cascade network on skin cancer dataset. In *8th International Conference on Computing in Engineering and Technology (ICCET 2023)*, volume 2023, pages 480–486, 2023. doi: 10.1049/icp.2023.1536.

- Jinfeng Wang, Qiming Huang, Feilong Tang, Jia Meng, Jionglong Su, and Sifan Song. Stepwise feature fusion: Local guides global. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022.
- Shansong Wang, Mojtaba Safari, Mingzhe Hu, Qiang Li, Chih-Wei Chang, Richard LJ Qiu, and Xiaofeng Yang. Dinov3 with test-time training for medical image registration, 2025. URL <https://arxiv.org/abs/2508.14809>.
- Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *ICCV*, 2021.
- Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024.
- Xinyu Xiong, Zihuang Wu, Lei Lu, and Yufa Xia. Sam3-unet: Simplified adaptation of segment anything model 3, 2025a. URL <https://arxiv.org/abs/2512.01789>.
- Xinyu Xiong, Zihuang Wu, Lei Zhang, Lei Lu, Ming Li, and Guanbin Li. Sam2-unext: an improved high-resolution baseline for adapting foundation models to downstream segmentation tasks. *arXiv preprint arXiv:2508.03566*, 2025b.
- Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *Visual Intelligence*, 4:2, 2026.
- Yimin Xu, Fan Yang, and Bin Xu. Dsu-net: An improved u-net model based on dinov2 and sam2 with multi-scale cross-model feature enhancement. In *2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 625–632, 2025.
- Zhiling Yan, Sifan Song, Dingjie Song, Yiwei Li, Rong Zhou, Weixiang Sun, Zhenhong Chen, Sekeun Kim, Hui Ren, Tianming Liu, et al. Samed-2: Selective memory enhanced medical segment anything model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*., 2025.
- Sicheng Yang, Hongqiu Wang, Zhaohu Xing, Sixiang Chen, and Lei Zhu. Segdino: An efficient design for medical and natural image segmentation with dino-v3. *ArXiv*, abs/2509.00833, 2025. URL <https://api.semanticscholar.org/CorpusID:281081098>.
- Sida Yi, Yuesheng Zhu, and Guibo Luo. PAAN: Pyramid attention augmented network for polyp segmentation. In *Medical Imaging with Deep Learning*, 2024. URL <https://openreview.net/forum?id=sz9baxSuxF>.
- Chen Zan, Peng Chenxu, Guo W, Xie L, and Wang S. Uncertainty-guided transformer for brain tumor segmentation. *Med Biol Eng Comput*), 2023.

Shuang Zeng, Chee Hong Lee, Micky C Nnamdi, Wenqi Shi, J Ben Tamo, Lei Zhu, Hangzhou He, Xinliang Zhang, Qian Chen, May D. Wang, Yanye Lu, and Qiushi Ren. Novel extraction of discriminative fine-grained feature to improve retinal vessel segmentation, 2025. URL <https://arxiv.org/abs/2505.03896>.

Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. 2018.

## Appendix A. Experimental Detail

### A.1. Loss Function

The loss function is designed to address key challenges in medical image segmentation, including boundary ambiguity and background interference. We employ a weighted combination of three components:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{BCE}} + \beta \mathcal{L}_{\text{Dice}} + \gamma \mathcal{L}_{\text{BL}} \quad (10)$$

where  $\alpha = 0.5$ ,  $\beta = 0.3$ ,  $\gamma = 0.2$ . The individual losses are defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (11)$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^N y_i p_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N p_i} \quad (12)$$

$$\mathcal{L}_{\text{BL}} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{y_i=1\}} \cdot (1 - p_i)^2 \quad (13)$$

#### Rationale for design:

- *BCE Loss* ( $\mathcal{L}_{\text{BCE}}$ ): Handles class imbalance by assigning higher weights to minority classes (polyps), crucial for medical images where polyp regions are small.
- *Dice Loss* ( $\mathcal{L}_{\text{Dice}}$ ): Directly optimizes the overlap between prediction and ground truth, critical for boundary precision in polyp segmentation.
- *Boundary Loss* ( $\mathcal{L}_{\text{BL}}$ ): Focuses on boundary pixels (where  $y_i = 1$ ) to improve edge delineation.

This combination effectively balances class imbalance, boundary accuracy, and overall segmentation quality.

## A.2. Experimental Setup

The experimental configuration is summarized in Table 11. The evaluation metrics include Dice, IoU, MAE,  $F_{w\beta}$ ,  $S_\alpha$ , F1-score, Acc and  $E_\phi$  as detailed in Section A.4. All experiments are conducted using the same training and testing protocols to ensure a fair comparison.

Table 11: Training and evaluation configuration.

Parameter	Value
Input resolution	$512 \times 512$
Batch size	8
Optimizer	AdamW ( $\eta = 0.0001$ , momentum=0.9)
Learning rate schedule	Step decay (0.1 at 50 epochs, 0.01 at 100 epochs)
Weight decay	$10^{-4}$
Epochs	100
Data augmentation	Horizontal flip, rotation, scaling, Zoom in, Zoom out, Brightening
Hardware	NVIDIA RTX 4090 GPU

## A.3. Datasets

Table 12: Dataset statistics (full URLs included).

Dataset	# Images	Modality	Source
Kvasir-SEG	1000	Endoscopic	<a href="https://github.com/DebeshJha/Kvasir-SEG">https://github.com/DebeshJha/Kvasir-SEG</a>
CVC-ClinicDB	612	Endoscopic	<a href="https://www.kaggle.com/datasets/balraj98/cvcclinicdb">https://www.kaggle.com/datasets/balraj98/cvcclinicdb</a>
CVC-ColonDB	380	Endoscopic	<a href="https://www.kaggle.com/datasets/longvil/cvc-colondb">https://www.kaggle.com/datasets/longvil/cvc-colondb</a>
CVC-300	60	Endoscopic	<a href="https://github.com/Polypproject/polyp_dataset">https://github.com/Polypproject/polyp_dataset</a>
DRIVE	40	Color Fundus	<a href="https://github.com/hmoghimifam/DRIVE">https://github.com/hmoghimifam/DRIVE</a>
STARE	20	Color Fundus	<a href="https://cecas.clemson.edu/~ahoover/stare/probing/index.html">https://cecas.clemson.edu/~ahoover/stare/probing/index.html</a>
Synapse	30 (3779 slices)	Abdominal CT	<a href="https://github.com/Always70/Synapse">https://github.com/Always70/Synapse</a>
MoNuSeg	30 (Train)	Histopathology	<a href="https://www.kaggle.com/datasets/tuanledinh/monuseg2018">https://www.kaggle.com/datasets/tuanledinh/monuseg2018</a>
CryoNuSeg	10 (per organ)	Histopathology	<a href="https://github.com/masih4/CryoNuSeg">https://github.com/masih4/CryoNuSeg</a>

We evaluate D<sup>2</sup>-Former on multiple medical image segmentation datasets as described below:

- **Kvasir-SEG**: Contains 1000 endoscopic images of polyps with pixel-wise annotations. Images exhibit varying polyp shapes, textures, and lighting conditions, making it challenging for boundary delineation.

- **CVC-ClinicDB**: A clinical dataset with 612 high-resolution colonoscopy images. Each image contains polyps with irregular shapes and blurred boundaries, along with normal tissues.
- **CVC-ColonDB**: Comprises 380 colonoscopy images with multiple polyps per image. Known for challenging visual conditions including poor contrast and complex backgrounds.
- **CVC-300**: Comprises 60 colonoscopy images with annotated polyp regions. This dataset is commonly used for polyp detection and segmentation tasks, featuring diverse visual conditions such as varying illumination, motion blur, and complex backgrounds.
- **DRIVE & STARE**: Widely used public benchmarks for retinal vessel segmentation. The DRIVE dataset contains 40 color fundus photographs with manual segmentation masks (Staal et al., 2004). The STARE dataset comprises 20 fundus images, also with vessel annotations, and is known for including pathological cases (Hoover et al., 2000). Both datasets are standard for evaluating algorithms’ ability to segment thin and complex vascular networks.
- **Synapse**: A public multi-organ abdominal CT dataset from the MICCAI 2015 challenge. It contains 30 contrast-enhanced abdominal CT scans (comprising 3779 axial slices in total) with pixel-level annotations for 8 abdominal organs (e.g., aorta, gallbladder, pancreas). It serves as a standard benchmark for evaluating 3D medical image segmentation models, particularly on handling challenges such as inter-organ variability, ambiguous boundaries, and large slice-wise variations (Chen et al., 2021b).
- **MoNuSeg**: A public challenge dataset for multi-organ nuclei instance segmentation in digital pathology. The training set includes 30 H&E-stained tissue images from 7 organs, with annotations for over 21,000 nuclei. It serves as a key benchmark for evaluating the generalization of nuclei segmentation algorithms across different tissue types (Kumar et al., 2020).
- **CryoNuSeg**: The first fully annotated dataset for nuclei instance segmentation in cryosectioned H&E-stained histological images. It contains images from 10 human organs and is specifically designed to address the challenges posed by frozen section samples, which differ in appearance from the more common formalin-fixed paraffin-embedded (FFPE) samples (Mahbod et al., 2021).

**Data Processing:** All datasets are preprocessed to  $512 \times 512$  resolution. For polyp segmentation, we use 1450 images (550 from CVC-ClinicDB and 900 from Kvasir) for training and 62 from CVC-ClinicDB, 100 from Kvasir, 62 from CVC-300 and 380 images from CVC-ColonDB for testing. Standard data augmentation (horizontal flip, rotation, and scaling) is applied to all training images to enhance model generalization. For the additional segmentation tasks, we adhered to the standard training and testing splits defined by the respective benchmark datasets to ensure a fair comparison. Specifically, for DRIVE and STARE, we used the standard split (20 train/20 test) and the common practice of training

on DRIVE and testing on STARE for cross-dataset evaluation. For Synapse, MoNuSeg, and CryoNuSeg, we followed the official challenge protocols and data splits.

#### A.4. Evaluation Metrics

We employ standard metrics to comprehensively assess segmentation performance:

- **Dice (Mean Dice Coefficient):**

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|}$$

Measures the overlap between predicted ( $X$ ) and ground truth ( $Y$ ) masks. Higher values indicate better segmentation accuracy, particularly for boundary precision.

- **mIoU (Mean Intersection over Union):**

$$\text{mIoU} = \frac{|X \cap Y|}{|X \cup Y|}$$

Evaluates the ratio of intersection to union of predicted and ground truth masks. A standard metric for segmentation tasks, with higher values indicating better overlap.

- **F<sub>β</sub> (F-measure):**

$$F_{\beta} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \quad (\beta^2 = 0.5)$$

Balances precision and recall with a bias toward recall ( $\beta^2 < 1$ ) to emphasize detection of small structures. Higher values indicate better trade-off between false positives and false negatives.

- **S<sub>α</sub> (Structure-measure):**

$$S_{\alpha} = \alpha \cdot SO + (1 - \alpha) \cdot SR \quad (\alpha = 0.06)$$

Combines object similarity ( $SO$ ) and region similarity ( $SR$ ) to evaluate structural consistency. Higher values indicate better preservation of anatomical structures.

- **E<sub>max</sub> (Enhanced-measure):**

$$E_{max} = \frac{(1 + \phi^2) \cdot SO \cdot SR}{\phi^2 \cdot SO + SR} \quad (\phi^2 = 15)$$

Refines the S<sub>α</sub> metric by emphasizing region coverage. Higher values indicate more complete segmentation of target regions.

- **MAE (Mean Absolute Error):**

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |p_i - y_i|$$

Quantifies the average absolute difference between predicted ( $p_i$ ) and ground truth ( $y_i$ ) pixels. Lower values indicate higher pixel-level accuracy, particularly important for boundary delineation.

- **F1-Score (F-Measure):**

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Balances precision and recall, providing a harmonic mean that is especially useful when class distribution is imbalanced, as often seen in vessel segmentation tasks.

- **Accuracy (Acc):**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Measures the overall proportion of correctly classified pixels (both vessel and background), offering a global assessment of segmentation correctness.

- **Aggregated Jaccard Index (AJI):**

$$AJI = \frac{\sum_i |G_i \cap P_{m(i)}|}{\sum_i |G_i \cup P_{m(i)}| + \sum_{k \in U} |P_k|}$$

Measures instance-level segmentation quality by penalizing over- and under-segmentation, where  $G_i$  denotes a ground-truth instance,  $P_{m(i)}$  its matched prediction, and  $U$  the set of unmatched predicted instances. Higher values indicate better instance delineation.

- **Panoptic Quality (PQ):**

$$PQ = \frac{\sum_{(p,g) \in TP} \text{IoU}(p, g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

Evaluates both segmentation quality and detection accuracy for instance segmentation, combining recognition and mask quality. Higher values indicate better overall instance segmentation performance.

- **Specificity (SP):**

$$SP = \frac{TN}{TN + FP}$$

Measures the proportion of correctly identified background pixels, reflecting the model's ability to suppress false positives. Higher values indicate better background discrimination.

- **Sensitivity (SE):**

$$SE = \frac{TP}{TP + FN}$$

Measures the proportion of correctly detected foreground pixels, reflecting the model's ability to capture target structures. Higher values indicate better foreground recall.

These metrics collectively assess accuracy, boundary precision, structural similarity, and robustness to class imbalance, providing a comprehensive evaluation of segmentation performance across diverse medical imaging tasks.

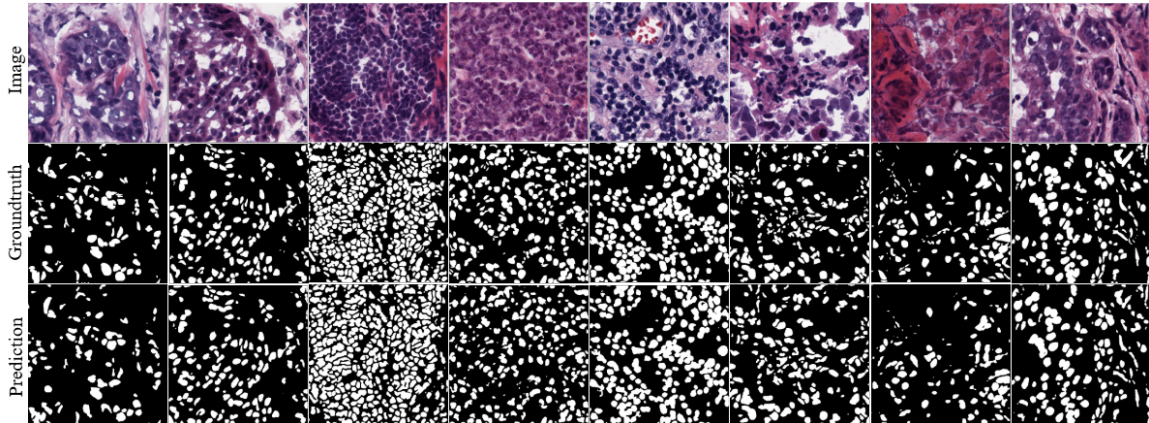


Figure 4: Additional qualitative results on nuclei image segmentation.

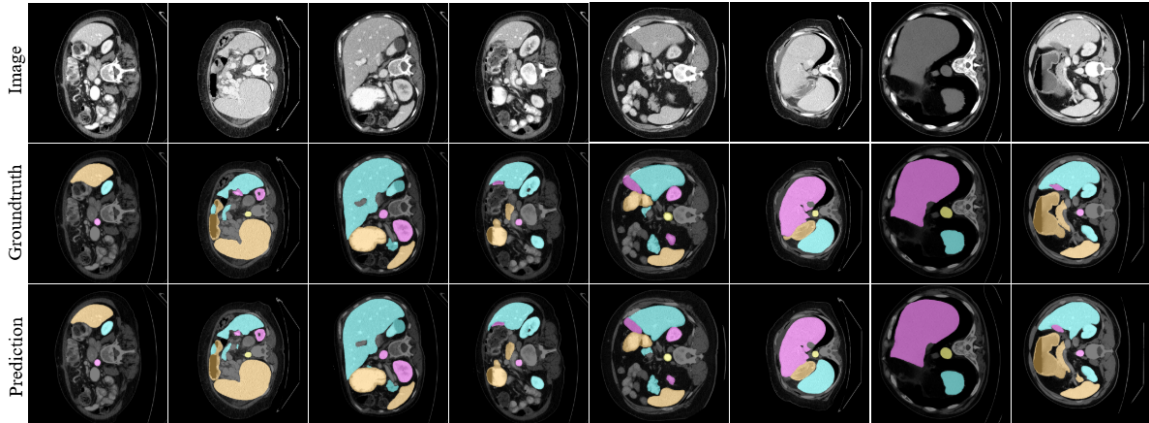


Figure 5: Additional qualitative results on Synapse image segmentation.

## Appendix B. Additional Results

### B.1. Appendix: Detailed Polyp Segmentation Results

Comprehensive quantitative results for polyp segmentation across four standard benchmarks are presented in Tables 11 and 12. Beyond the Dice and mIoU scores reported in the main text, our D<sup>2</sup>-Former framework achieves superior or highly competitive performance on multiple additional metrics, including the weighted F-measure ( $F_{sp}$ ), structure-measure ( $S_\alpha$ ), enhanced-alignment measure ( $E_p^{max}$ ), and mean absolute error (MAE).

Specifically, on the seen-domain datasets, Kvasir-SEG and CVC-ClinicDB, D<sup>2</sup>-Former attains the highest  $F_{sp}$  scores of 94.2 and 94.3, respectively, and competitive scores on  $S_\alpha$  and  $E_p^{max}$ . On the more challenging unseen-domain datasets, CVC-300 and CVC-ColonDB, our model demonstrates strong generalization, leading in  $F_{sp}$  (93.8 on CVC-300, 85.6 on CVC-ColonDB) while achieving the best  $S_\alpha$  score (95.5) on CVC-300. The results across

Table 13: Quantitative results comparison of D<sup>2</sup>-Former with SOTA methods on Kvasir-SEG and CVC-ClinicDB datasets.

Methods	Kvasir-SEG						CVC-ClinicDB					
	mDice $\uparrow$	mIoU $\uparrow$	$F_{sp}$ $\uparrow$	$S_{\alpha}$ $\uparrow$	$E_p^{max}$ $\uparrow$	MAE $\downarrow$	mDice $\uparrow$	mIoU $\uparrow$	$F_{sp}$ $\uparrow$	$S_{\alpha}$ $\uparrow$	$E_p^{max}$ $\uparrow$	MAE $\downarrow$
U-Net	81.8	74.6	79.4	85.8	89.3	5.5	82.3	75.5	81.1	88.9	95.4	1.9
PraNet	89.8	84.0	88.5	91.5	94.8	3.0	89.9	84.9	89.6	93.6	97.9	0.9
Polyp-PVT	91.7	86.4	91.1	92.5	95.6	2.3	93.7	88.9	93.6	94.9	98.5	0.6
VM-UNet	91.3	85.6	90.2	91.8	95.8	2.7	92.6	87.1	92.7	93.3	97.1	0.9
CoinNet	92.6	87.2	93.9	92.6	97.9	2.0	93.0	88.7	94.0	95.2	98.7	0.6
D <sup>2</sup> -Former	93.4	87.3	94.2	93.6	96.1	2.2	94.2	88.9	94.3	95.5	98.2	0.6

Table 14: Quantitative results comparison of D<sup>2</sup>-Former with SOTA methods on Unseen CVC-300 and CVC-ColonDB datasets.

Methods	CVC-300						CVC-ColonDB					
	mDice $\uparrow$	mIoU $\uparrow$	$F_{sp}$ $\uparrow$	$S_{\alpha}$ $\uparrow$	$E_p^{max}$ $\uparrow$	MAE $\downarrow$	mDice $\uparrow$	mIoU $\uparrow$	$F_{sp}$ $\uparrow$	$S_{\alpha}$ $\uparrow$	$E_p^{max}$ $\uparrow$	MAE $\downarrow$
U-Net	71.0	62.7	68.4	84.3	87.6	2.2	51.2	44.4	49.8	71.2	77.6	6.1
PraNet	87.1	79.7	84.3	92.5	97.2	1.0	70.9	64.0	69.6	81.9	86.9	4.5
Polyp-PVT	90.0	83.3	88.4	93.5	97.3	0.7	80.8	72.7	79.5	86.5	91.3	3.1
VM-UNet	88.6	81.8	84.9	92.1	96.8	0.9	79.8	71.2	78.2	86.1	90.4	3.6
CoinNet	90.9	86.3	88.1	94.2	98.9	0.5	79.7	72.9	78.9	87.5	89.7	2.2
D <sup>2</sup> -Former	92.0	86.1	93.8	95.5	98.1	0.6	82.8	76.3	85.6	89.8	93.3	1.9

this broader set of evaluation metrics consistently reinforce the robustness and precision of the proposed model for polyp segmentation across diverse clinical scenarios.

## B.2. Appendix: Additional results from Synapse and nuclei datasets

Figure 5 and Figure 4 presents qualitative segmentation results on the Synapse multi-organ CT and nuclei datasets. The visualizations demonstrate that D<sup>2</sup>-Former accurately delineates complex organ boundaries in abdominal CT scans, including challenging structures like the pancreas and gallbladder, and effectively segments individual nuclei in dense histopathology images. These results highlight the model’s strong generalization across both macro-scale anatomical and micro-scale cellular segmentation tasks.

## Appendix C. Responsible AI and Deployment Considerations

Although we evaluate on multiple public benchmarks, these datasets may not fully represent the diversity of real clinical settings (e.g., scanner protocols, institutions, patient populations, and rare pathologies), and annotations can contain inter-observer variability. As a result, performance may degrade under domain shift and may exhibit systematic errors for under-represented cases. In deployment, segmentation errors (e.g., over-segmentation that inflates lesion extent or false positives on normal tissue) could contribute to unnecessary follow-up or intervention if used without clinical oversight. We therefore emphasize that D<sup>2</sup>-Former is intended as decision support; prospective validation, site-specific calibration/thresholding, and human-in-the-loop review are necessary before any clinical use.