

MedCAGD: Context-Aware Gated Decoder for Efficient Medical Image Segmentation

Saad Wazir¹, Patrick Dominique Vibild², Dinh Phu Tran¹, Seongah Kim¹, and Daeyoung Kim¹

¹ School of Computing, Korea Advanced Institute of Science and Technology (KAIST),
Daejeon, Republic of Korea

{saad.wazir, phutx2000, kimsa0322, kimd}@kaist.ac.kr

² Department of Energy, Aalborg University, Aalborg, Denmark
padovi@energy.aau.dk

Abstract. Medical image segmentation relies on the ability of encoder-decoder architectures to translate rich feature representations into accurate pixel-level predictions under challenging conditions such as low contrast, structural ambiguity, and scale variability. While recent advances in large-scale pretraining and transformer-based encoders have substantially improved feature extraction, segmentation accuracy remains constrained by decoder design, particularly in terms of cross-scale alignment, contextual integration, and boundary preservation. In this work, we revisit medical image segmentation from a decoder-centric perspective and propose a context-aware gated decoder that systematically regulates feature fusion and contextual aggregation throughout the decoding process. The proposed decoder integrates lightweight multi-scale channel recalibration, gated skip fusion with spatial competition and a global context aggregation mechanism that injects encoder-wide information into intermediate decoding stages. This design enables effective translation of strong pretrained encoder representations into spatially consistent predictions. Extensive experiments across 11 medical image segmentation benchmarks validate the effectiveness and demonstrate that the proposed approach consistently outperforms strong baselines while remaining computationally practical. Code: <https://github.com/saadwazir/MedCAGD>

Keywords: Medical Image Segmentation · Decoder Design · Bio-informatics

1 Introduction

Medical image segmentation is fundamental for quantitative analysis, diagnosis, treatment planning, and clinical assessment. Tasks such as organ delineation, lesion localization, tumor boundary extraction, and cellular segmentation require pixel-level precision under challenging conditions. To address these challenges, encoder-decoder architectures, particularly U-Net [40] based models, have become the dominant paradigm in medical image segmentation. Within this framework, performance is increasingly governed by encoder-decoder design, skip-connection formulation, feature fusion, and, more prominently, improvements in encoder capacity [38, 39, 44]. Although attention mechanisms have evolved from local CNN-based modules to non-local and transformer-based formulations for long-range context modeling [34, 61], their computational cost

and design limitations leave the integration of global context during decoding unresolved. Consequently, many segmentation errors arise from suboptimal decoding and cross-scale alignment rather than insufficient feature extraction [13, 37].

Recently, foundation model approaches have demonstrated strong cross-domain generalization in vision tasks. In segmentation, SAM [24] has introduced a promptable, generalist paradigm, inspiring SAM-derived medical variants [3, 42, 60, 64] that demonstrate strong generalization. However, even medically adapted versions require substantial labeled data, modality specific supervision, and significant computational resources to approach the performance of specialist medical segmentation models [28].

In parallel, advances in large-scale pretraining have strengthened encoder representations [27, 46, 53, 54, 58], leading modern segmentation frameworks to adopt powerful pretrained encoders [12, 36, 38], yet accuracy remains dependent on effective decoding into spatially consistent predictions, with boundary errors and fragmentation often stemming from semantic misalignment rather than weak representations [37, 57]. Taken together, these observations indicate that continued improvements in medical image segmentation accuracy increasingly hinge on decoder design rather than encoder capacity alone. In this context, this work explores decoder design as a complementary and computationally efficient approach for improving segmentation accuracy. Designing such decoders remains challenging, as they must balance contextual integration with spatial precision without incurring excessive computational cost. We argue that meaningful performance gains can be achieved through a principled decoder centric framework that systematically regulates contextual aggregation across decoding stages, enabling more faithful translation of strong encoder representations into accurate pixel-level predictions without increasing encoder complexity or relying on task specific fine tuning. Our main contributions are:

- **MedCAGD: Context-Aware Gated Decoder Architecture.** We propose a decoder centric segmentation framework that systematically regulates feature transformation during decoding. The architecture integrates Bottleneck with Global Context Injection, Spatially Competitive Attention Gate based skip regulation, Multi-level Context aggregation, and stage wise refinement, positioning decoder design as the primary factor governing accurate pixel-level prediction.
- **Structured Context Regulated Decoder Components.** We introduce a unified set of modules that directly correspond to the methodological components of MedCAGD: (i) **Efficient Channel Attention Block with multi-scale Pooling**, which performs context sensitive channel recalibration using multi-scale descriptors and normalized channel competition. (ii) **Spatially Competitive Attention Gate**, which formulates skip fusion as normalized multiplicative encoder-decoder agreement combined with global modulation and multi-scale spatial competition. (iii) **Multi-level Context Aggregation with Residual Attention**, which injects globally coherent multi-level encoder semantics into intermediate decoder stages to mitigate cross scale semantic misalignment. (iv) **Refinement Block** with local refinement and channel recalibration, which strengthens local reconstruction and stabilizes feature propagation across decoding stages.
- **Encoder agnostic and computationally efficient design with strong empirical validation.** The proposed MedCAGD remains fully encoder-agnostic through Uni-

versal Feature Projection, enabling broad compatibility with PyTorch timm encoders, while maintaining practical complexity of **30.60 M parameters** and **5.0 GFLOPs**. Extensive experiments across 11 heterogeneous medical image segmentation benchmarks demonstrate consistent improvements over strong CNN, Transformer, Mamba, SAM, and recent decoder centric baselines.

2 Related Work

CNNs have been the cornerstone of medical image segmentation, most notably U-Net [40], which became dominant by combining hierarchical features with skip connections to recover fine spatial detail. Building on this design, a wide range of U-Net variants [18, 35, 55, 56, 66] emerged. These works introduced dense skip connections, nested U-Net designs, and multi-scale aggregation for improved context and boundaries, while nnU-Net [20] highlighted the role of systematic pipeline optimization. However, CNNs still rely on local operations, limiting long-range dependency modeling. **Attention mechanisms** [32, 33, 65] partially mitigate this by enhancing features via channel, spatial, and residual attention, but mainly recalibrate features without modeling global interactions.

Transformer based architectures address the limitation of long-range dependency modeling by introducing self-attention. TransUNet [6] pioneered the integration of Vision Transformers with convolutional decoders. Subsequent architectures such as Swin-UNet [4] adopted hierarchical Transformer designs with shifted window attention to improve computational efficiency. Some architectures such as UNeXt [47] replace self-attention with convolutional MLP-based designs to reduce computational overhead, while task-specific models such as PraNet [12] introduce structurally motivated attention mechanisms to enhance boundary cues without relying on full global attention. However, Transformers suffer from quadratic computational and memory complexity, limiting scalability.

Mamba [14] addresses the quadratic computational and memory inefficiency of Transformers by replacing explicit attention with linear time state space modeling. Several recent works have explored Mamba-based architectures for medical image segmentation. VM-UNet [41] introduces Vision Mamba blocks into a U-Net style architecture to enhance long-range spatial dependency modeling while maintaining linear computational complexity. U-Mamba [29] further integrates Mamba blocks into CNN encoders within the nnU-Net framework, combining local convolutional feature extraction with state space modeling to improve global context representation. Swin-UMamba [26] extends this by incorporating hierarchical representations and ImageNet pretrained Mamba-based encoders. Existing Mamba-based segmentation methods primarily emphasize encoder representations and typically operate with a fixed state size, which may limit performance scalability across tasks of varying complexity.

Decoder design has been advanced through multi-scale context aggregation [62], dense [66] or full scale skip connections [18], deep supervision [5, 63], efficient spatial reconstruction modules [38], dual decoder architectures [45], and the integration of transformer blocks [4]. UCTransNet [49] replaces fixed skip connections with learnable semantics aware fusion to better align multi-scale features while preserving spa-

tial detail, while PolypPVT [10] embed CBAM [59] within the decoding stage for enhanced feature refinement. MCADS [57] follows a complementary direction inspired by [32, 51], combining residual linear attention with depth to space based upsampling to preserve fine structural details during resolution recovery, achieving higher accuracy at the expense of efficiency. More recently, EMCAD [37] introduces a convolutional decoder that integrates multi-stage hybrid transformer encoder features using modified and enhanced attention mechanisms following [17, 33, 59], leading to strong performance in medical image segmentation. Despite recent progress, segmentation performance remains fundamentally constrained by decoder design. In particular, challenges in cross-scale alignment, boundary refinement, and long-range context translation continue to persist across the previously discussed architectures. Although these methods introduce increasingly sophisticated attention, upsampling, and feature fusion strategies, they often prioritize stronger pretrained encoders while relying on decoding mechanisms that inadequately preserve fine spatial detail and global semantic consistency. Consequently, they achieve only modest improvements and inconsistent performance across tasks of varying anatomical complexity and structural variability. Collectively, these observations indicate that the primary bottleneck lies in decoder formulation rather than encoder capacity alone, motivating the exploration of more principled and context-aware decoder designs for medical image segmentation.

Foundation models, particularly **Segment Anything Model (SAM)** [24] has demonstrated strong generalization across diverse image segmentation tasks through prompt driven interaction, enabling flexible mask generation. In the medical imaging domain, several adaptations such as AutoSAM [42], Medical SAM3 [22], SAMed [64], SAM3D [3] and Self-Prompt-SAM [60] have explored fine tuning strategies, adapter based training, and learned prompting mechanisms to better align SAM with domain specific structures. While these approaches improve robustness, they typically rely on explicit prompting, large curated datasets for adaptation, and substantial computational resources. Empirical studies further indicate that SAM based methods often under-perform specialist architectures on fixed task, particularly for datasets characterized by subtle boundaries or fine grained anatomical structures, such as fundus imaging [28]. Although not the primary focus of this study, SAM based methods are included to enrich the analysis and provide a broader contextual understanding.

3 Methodology

In this section, we first present the overall encoder–decoder architecture and explain how its components are integrated to regulate feature flow during decoding. We then describe the fundamental modules that form the foundation of the proposed method, along with a brief introduction to the encoder. Finally, we introduce the training objective. The complete pipeline of the proposed approach is depicted in Fig. 1.

3.1 Overall Decoder Architecture and Component Integration

Multi-scale hierarchical features are first extracted by the encoder and projected into fixed dimensional representations. Decoding begins from the bottleneck output and pro-

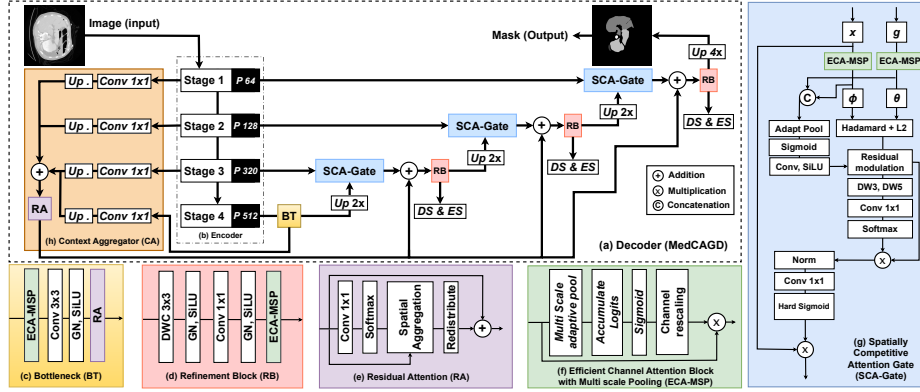


Fig. 1: Overview of (a) MedCAGD, the proposed decoder architecture. (b) Multi-scale encoder features are projected into a unified decoder feature space. (c) Bottleneck (BT) initializes decoding by refining the deepest encoder feature using (f) Efficient Channel Attention with Multi-scale Pooling (ECA-MSP) for adaptive channel recalibration and (e) Residual Attention (RA) for global context integration. (g) Spatially Competitive Attention Gate (SCA-Gate) selectively regulates encoder skip features before fusion with decoder features. (h) Context Aggregator (CA) injects globally aggregated multi-scale semantics into each decoding stage. (d) Refinement Block (RB) enhances fused decoder features through efficient local refinement and channel recalibration. Deep Supervision (DS) and Edge Supervision (ES) provide auxiliary supervision during training.

ceeds through a sequence of decoder blocks with progressively increasing spatial resolution. At each stage, the current decoder feature is first upsampled and then fused with the corresponding encoder feature through the Spatially Competitive Attention Gate, enabling selective and context-aware regulation of encoder features prior to concatenation. In parallel, Multi-level Context Aggregation operates on the projected encoder features, and the resulting context representation is added residually to the decoder feature before refinement. This ensures that each stage is guided by globally aggregated multi-scale semantics while preserving stage-specific reconstruction. The updated feature is then passed through the Refinement Block for convolutional enhancement and channel recalibration. This sequence of upsampling, gated skip fusion, context aggregation, and refinement is repeated across decoding stages. At the final stage, the full resolution decoder feature is forwarded to the segmentation head to produce the primary prediction. Intermediate decoder features are additionally connected to auxiliary segmentation and edge prediction heads to enable deep supervision during training.

3.2 Encoder and Universal Feature Projection

We employ an ImageNet pretrained PVTv2-B2 as the encoder due to its hierarchical transformer design, which provides multi-level features well aligned with our decoder. It achieves a strong balance between accuracy and efficiency, as validated in Sec. 5.3. Furthermore, its adoption by several SOTA decoder centric methods ensures fair comparison. As shown in Fig. 1 (b), given multi-scale encoder feature maps $\{c_i\}_{i=1}^4$, each

encoder feature c_i is aligned to the predefined decoder channel dimension using a learnable 1×1 convolutional projection $\mathcal{P}_i(\cdot)$, such that $p_i = \mathcal{P}_i(c_i)$, where the decoder channels are fixed to 64, 128, 320, and 512 across stages. This projection ensures consistent decoder dimensionality while preserving the multi-scale hierarchy of the encoder.

3.3 Residual Attention (RA)

To model global spatial dependencies during decoding, we employ a lightweight non-local attention mechanism embedded within a residual formulation. As shown in Fig. 1 (e), given an input feature map X , a spatial importance distribution is first computed using a pointwise projection followed by softmax normalization. This distribution is used to aggregate long-range spatial responses into a global context descriptor. The aggregated vector is then transformed through a lightweight channel mixing function with intermediate dimensionality reduction and reinjected into the feature stream via residual addition, yielding

$$Y = X + \mathcal{P}_2 \left(\delta \left(\mathcal{P}_1 \left(\sum_{i=1}^{HW} \text{Softmax}(\mathcal{P}_0(X))_i X_i \right) \right) \right). \quad (1)$$

Here, $\mathcal{P}_0(\cdot)$, $\mathcal{P}_1(\cdot)$, and $\mathcal{P}_2(\cdot)$ denote learnable pointwise convolutional projections, and $\delta(\cdot)$ denotes a nonlinear activation. The residual formulation preserves local structure while enabling efficient global context integration, as reported in Sec. 5.1, where enabling RA improves performance.

3.4 Efficient Channel Attention Block with Multi-scale Pooling (ECA-MSP)

To adaptively recalibrate channel responses based on contextual relevance, we employ an Efficient Channel Attention block extended with multi-scale pooling. As shown in Fig. 1(f), given an input feature map X , channel descriptors are extracted at multiple contextual granularities using adaptive average pooling. Here, the pooling scale refers to the target spatial resolution of adaptive average pooling used to compute channel statistics, while the pooling operation itself remains average pooling. For a set of pooling scales $\mathcal{S} = \{1, 2, 4\}$, multi-scale channel descriptors are independently transformed through a one-dimensional convolution that models local cross channel interaction without dimensionality reduction. The resulting responses are aggregated across scales and converted into channel attention weights. The overall operation is expressed as

$$X' = X \odot \sigma(A_{\text{ms}}(X)), \quad A_{\text{ms}}(X) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \psi(\text{AdaptiveAvgPool}_s(X)). \quad (2)$$

where $\text{AdaptiveAvgPool}_s(\cdot)$ denotes adaptive average pooling to spatial size $s \times s$ followed by spatial aggregation to obtain channel descriptors, $\psi(\cdot)$ denotes local cross channel interaction implemented via one dimensional convolution, and $\sigma(\cdot)$ denotes the sigmoid activation. Unlike SE-Net [17], ECA-Net [52], and EMCAD [37], which

rely on single scale global descriptors or bottleneck based dual pooling, the proposed formulation leverages multi-scale pooling to capture complementary contextual information, ranging from global semantic statistics to coarse localized cues. Sec. 5.1 and 5.2 provides empirical evidence of the effectiveness of ECA-MSP.

3.5 Bottleneck (BT) with Global Context Injection

At the deepest stage of the network, a bottleneck module refines the highest level encoder feature and injects global context before decoding begins as shown in Fig. 1 (c). Let F_4 denote the projected deepest encoder feature. Channel responses are first recalibrated using the ECA-MSP $\mathcal{E}(\cdot)$, followed by convolutional refinement $\rho(\cdot)$, and finally global context injection through the RA operator $\mathcal{R}(\cdot)$. The overall bottleneck transformation is expressed as $B = \mathcal{R}(\rho(\mathcal{E}(F_4)))$. This formulation enables the decoder to start from a context-aware semantic representation while preserving the structural properties of the refined feature map. Further validated by ablation study in Sec. 5.1.

3.6 Spatially Competitive Attention Gate (SCA-Gate)

Recent studies [49, 50] indicate that skip connections in encoder-decoder architectures are not universally beneficial, since indiscriminate feature propagation can introduce semantically incompatible information due to the encoder-decoder semantic gap. Following this motivation, we formulate skip connections as learnable and selective feature regulation mechanisms rather than passive information pathways, as shown in Fig. 1 (g). Let g denote the decoder feature at a given stage and x the corresponding encoder skip feature. Both features are first recalibrated using ECA-MSP $\mathcal{E}(\cdot)$ and projected into a shared latent space via lightweight transformations $\theta(\cdot)$ and $\phi(\cdot)$. Their interaction is modeled as $f = \theta(\mathcal{E}(g)) \odot \phi(\mathcal{E}(x))$. The gated skip feature is then defined as

$$\begin{aligned} x' &= x \odot \sigma(\mathcal{H}(f, g, x)), \\ \mathcal{H}(f, g, x) &= f \odot (1 + \mathcal{G}(g, x)) \odot (1 + \mathcal{S}(f)). \end{aligned} \quad (3)$$

Here, $\mathcal{G}(\cdot)$ denotes global channel modulation derived from the joint encoder-decoder representation, while $\mathcal{S}(\cdot)$ represents multi-scale spatial competition. In practice, $\mathcal{S}(f)$ is implemented using parallel depthwise convolutions with kernel sizes 3 and 5, namely $\mathcal{D}_3(f)$ and $\mathcal{D}_5(f)$, whose aggregated responses are normalized via temperature-controlled softmax. The resulting attention mask $\sigma(\mathcal{H}(f, g, x))$ is multiplicatively applied to the skip feature x . The function $\sigma(\cdot)$ denotes a bounded activation for adaptive skip regulation. Unlike Attention U-Net [33] and EMCAD [37], which rely on additive fusion followed by sigmoid masking, the proposed formulation models skip selection as normalized multiplicative agreement combined with global modulation and spatial competition. The effectiveness of SCA-Gate is further validated by comprehensive ablation studies in Sec. 5.1 and 5.2.

3.7 Context Aggregator (CA)

While skip connections transfer information between corresponding encoder and decoder stages, effective decoding also requires global awareness across multiple semantic scales. To this end, we introduce a multi-level context aggregation module, as

shown in Fig. 1(h), which integrates features from different levels and injects globally consistent contextual information into each decoding stage as supported by the ablation study in Sec. 5.1. Let $\{F_k\}_{k=1}^K$ denote feature maps from multiple encoder stages. Each feature is projected into a unified channel space and spatially aligned to the target decoder resolution using learnable pointwise projections $\mathcal{P}_k(\cdot)$ with interpolation. The aligned features are averaged and refined through the RA operator, producing $F_{\text{ctx}} = \mathcal{R}\left(\frac{1}{K} \sum_{k=1}^K \mathcal{P}_k(F_k)\right)$. The resulting representation aggregates globally coherent multi-scale semantics and is added residually to the decoder feature, providing stage independent global guidance that complements context-aware gated skip fusion.

3.8 Refinement Block (RB)

The Refinement Block, shown in Fig. 1(d), enhances decoder feature propagation through sequential local refinement and adaptive channel recalibration, as validated by the ablation study in Sec. 5.1. Given an input feature, it is first processed by a depthwise convolution for spatial filtering, followed by a pointwise convolution for channel mixing, each combined with Group Normalization and SiLU activation. This design enhances spatial consistency while maintaining computational efficiency during local refinement. The refined features are subsequently recalibrated using ECA-MSP.

3.9 Segmentation Outputs and Training Objective

The decoder produces the final segmentation along with auxiliary segmentation and edge predictions for supervision. The final feature generates the primary logit, while intermediate features are independently projected and upsampled to the input resolution for deep supervision. Let $\hat{Y}$ denote the final segmentation logit and $\{\hat{Y}_i\}_{i=1}^3$ the auxiliary segmentation logits. Deep supervision is applied by optimizing a weighted and normalized sum of losses over these predictions to promote consistent optimization across decoding depths. In parallel, auxiliary edge logits $\{\hat{E}_i\}_{i=1}^3$ are generated from intermediate decoder features and supervised using binary edge targets derived from the ground truth masks, encouraging boundary aware decoding. All segmentation and edge predictions are optimized using the binary cross entropy (BCE) loss [21]. The overall training objective combines the main segmentation loss with deep supervision and edge supervision losses using normalized weights. Ablation results in Fig. 3 demonstrate that enabling both deep supervision (DS) and edge supervision (ES) consistently improves Dice and HD95 across six segmentation benchmarks, highlighting their complementary roles in dense prediction and contour refinement. During inference, only the final segmentation prediction is retained.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluated the proposed method on 11 publicly available medical image segmentation datasets that have also been benchmarked in recent SOTA studies, including

Table 1: Comprehensive performance comparison across 9 medical image segmentation benchmarks. Average Dice scores $\uparrow$ are reported. Bold and underline denote the best and second best results, respectively. All methods were reproduced and averaged over five runs, with fine tuning applied to SOTA models for fair comparison. Results marked with * are reported from the papers. “-” indicates unavailable results.

Method	Params $\downarrow$	Flops $\downarrow$	Skin		Polyp		Fundus		Neoplasm		Cell	All
			ISIC17	ISIC18	ETIS	ColonDB	DRIVE	FIVES	BUSI	ThyroidXL	CellSeg	Avg
U-Net [40]	34.53 M	65.53 G	83.07	86.67	76.85	83.95	71.20	75.77	74.04	71.16	71.52	77.14
AttnUNet [33]	34.88 M	66.64 G	83.66	87.05	76.84	86.46	71.68	75.99	74.48	72.50	72.64	77.92
DeepLabv3+ [7]	39.76 M	14.92 G	83.84	88.64	90.73	91.92	69.59	75.12	76.81	73.46	71.90	80.22
UNet++ [66]	<u>09.16 M</u>	34.65 G	82.98	87.46	77.40	87.88	72.94	<u>85.74</u>	74.46	83.94	78.30	81.23
nnU-Net [20]	31.29 M	55.26 G	83.23	88.53	80.13	91.63	75.43	76.10	76.46	86.08	83.53	82.34
PraNet [12]	32.55 M	06.93 G	83.03	88.56	83.84	89.16	75.21	84.57	75.14	85.51	79.07	82.68
TransUNet [6]	105.32 M	38.52 G	85.00	89.16	87.79	91.63	74.98	83.54	78.30	85.77	79.08	83.92
Swin-Unet [4]	27.17 M	06.20 G	83.97	89.26	85.10	89.27	74.93	84.17	77.38	85.80	78.84	83.19
UCTransNet [49]	65.60 M	56.70 G	83.27	89.18	87.35	91.65	75.42	84.74	79.53	85.82	79.33	84.03
UNeXt [47]	1.470 M	0.570 G	82.74	87.78	74.03	83.84	74.77	76.60	74.71	84.46	75.71	79.40
VM-UNet [41]	27.43 M	<u>04.12 G</u>	<u>85.99</u>	87.05	85.52	88.71	73.25	83.51	74.69	78.31	74.94	81.33
Swin-U-Mamba [26]	60.00 M	68.00 G	83.40	87.62	86.63	87.97	73.32	82.66	73.38	84.96	75.56	81.72
EMCAD [37]	26.76 M	05.60 G	85.95	90.96	<u>92.29</u>	<u>92.31</u>	77.15	82.51	<u>80.25</u>	83.33	79.13	84.87
MCADS [57]	50.90 M	61.89 G	84.14	<u>91.01</u>	92.24	91.37	<u>78.42</u>	76.05	80.03	<u>86.33</u>	86.68	<u>85.14</u>
Ours	30.60 M	05.00 G	86.61	91.56	93.47	93.27	81.63	87.50	83.47	88.02	<u>86.61</u>	88.01
AutoSam [42]*	41.56 M	25.11 G	-	-	79.70	83.00	-	-	-	-	-	-
Medical SAM3 [22]*	840.0 M	-	-	-	86.10	-	55.80	-	-	-	-	-

Table 2: Performance comparison with SOTA methods on the Synapse multi-organ dataset. Overall Dice, IoU, and HD95 are reported together with per class Dice scores. All methods were reproduced and averaged over five runs, with fine tuning applied to SOTA models for fair comparison. Results marked with * are reported from the papers. “-” indicates unavailable results.

Method	Dice $\uparrow$	IoU $\uparrow$	HD95 $\downarrow$	Aorta	GB	KL	KR	Liver	PC	SP	SM
U-Net [40]	70.11	59.39	44.69	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet [33]	71.70	68.09	26.01	84.04	66.42	57.26	84.53	81.28	73.87	66.06	60.17
UNet++ [66]	72.39	68.82	25.61	83.65	67.66	57.26	84.53	81.34	73.87	68.97	61.85
nnU-Net [20]	75.33	71.47	19.34	77.06	73.27	76.34	84.53	79.98	73.34	77.62	60.52
PraNetV2 [16]	83.75	74.81	17.77	88.69	72.79	85.41	82.91	<u>95.82</u>	68.47	<u>93.09</u>	<u>85.85</u>
TransUNet [6]	77.61	67.32	26.90	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75.00
Swin-Unet [4]	77.58	66.88	27.32	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
UCTransNet [49]	79.08	75.41	15.59	83.06	81.35	77.24	78.23	85.76	74.77	81.89	70.31
UNETR [15]*	78.35	-	18.59	89.80	56.30	85.60	84.52	94.57	60.47	85.00	70.46
MISSFormer [19]*	81.96	-	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
U-Mamba [29]	78.63	74.87	16.19	83.77	78.70	79.40	82.37	83.86	74.78	79.77	66.41
VM-UNet [41]	73.39	71.61	27.97	63.57	72.62	77.98	92.59	79.44	70.80	55.58	74.55
EMCAD [37]	83.63	74.65	15.68	88.14	68.87	<u>88.08</u>	84.10	95.26	68.51	92.17	83.92
MCADS [57]	85.03	<u>81.71</u>	11.11	90.81	<u>86.07</u>	86.77	83.24	87.66	83.55	85.74	76.38
Ours	87.00 $\pm$ 0.2	83.77	<u>14.39</u>	92.28	90.31	89.72	<u>87.21</u>	91.02	<u>82.08</u>	86.91	76.51
Self -Prompt SAM [60]*	<u>86.74</u>	-	-	<u>91.99</u>	69.95	85.65	85.40	97.39	79.18	94.38	89.94

EMCAD (CVPR 2024) [37], Swin-UMamba (MICCAI 2024) [26], ThyroidXL (MICCAI 2025) [11], and Medical-SAM3 [22] (2026). The datasets span diverse organs and imaging modalities, including ISIC17 [9], ISIC18 [8], ETIS [48], ColonDB [48], DRIVE [43], FIVES [23], BUSI [1], ThyroidXL [11], CellSeg [30], Synapse [25], and ACDC [2], covering dermoscopy, endoscopy, fundus imaging, ultrasound, microscopy, CT, and MRI. Performance was assessed using Dice, IoU, and HD95 [31]. Additional dataset and metric details are provided in the supplementary material.

4.2 Implementation details

We implemented the proposed network in PyTorch 2.7 and conducted all experiments on a single NVIDIA RTX 3090 GPU with 24 GB of memory. The model was optimized using AdamW with a learning rate of $1e-4$. Training was performed for over 300 epochs with a batch size of 16, and the best model was selected based on the validation Dice score. All input images were resized to 224 by 224, and online data augmentation including random rotation, horizontal and vertical flipping, and random cropping was applied. The network was trained using a BCE loss. For the DRIVE and FIVES datasets, we generated 256 by 256 overlapping patches with a stride of 128 for training. For CellSeg, we generated 384 by 384 overlapping patches with a stride of 192 for training. During testing, similar overlapping patches were extracted, predictions were obtained for each patch, and the full resolution segmentation maps were reconstructed for evaluation. To ensure a fair comparison, all competing methods were reproduced using their publicly available implementations, and the results were averaged over five independent runs.

4.3 Results

We compare our method with representative CNN, transformer, Mamba, SAM, and decoder centric models on 2D binary and multi-class benchmarks. Across the 9 binary-class segmentation datasets in Table 1, our method consistently outperforms all baselines on skin, polyp, fundus, neoplasm, and cell tasks. Conventional CNNs remain stable but are constrained by limited global modeling, while nnU-Net improves performance through optimization without closing the performance gap. Transformer models strengthen global reasoning, with UCTransNet achieving competitive results at the cost of higher architectural complexity. Mamba variants model long-range dependencies efficiently yet yield only marginal or unstable gains. Decoder centric approaches, especially EMCAD and MCADS, form the strongest baselines, underscoring the importance of feature fusion. However, our results demonstrate that structured context-aware skip gating yields superior performance without relying on larger or heavier designs. SAM based foundation models underperform on domain specific medical data, highlighting the necessity of task tailored architectures. On the Synapse multi-class segmentation dataset in Table 2, our method achieves the highest overall Dice and IoU while remaining competitive in boundary accuracy. CNN baselines struggle with small and complex organs, transformers improve structural coherence but exhibit class level variability, and Mamba models do not consistently minimize boundary errors. Although MCADS

demonstrates strong boundary performance, our method maintains a better overall balance between accuracy and structural consistency. On the ACDC multi-class segmentation dataset in Table 3, our method achieves the best Dice and the lowest HD95, clearly surpassing CNN, transformer, and Mamba models. The substantial reduction in HD95 reflects sharper boundary delineation across RV, Myo, and LV, confirming that context-aware gated fusion enhances both anatomical coherence and fine structural precision. Overall, simply increasing encoder scale or global modeling capacity is insufficient. Effective decoder design is decisive, as reflected by consistently superior performance over strong baselines.

In terms of **computational cost**, our method maintains a strong efficiency profile, as evident in Table 1, while delivering superior segmentation accuracy. Compared with SOTA decoder centric approaches EMCAD and MCADS, our design achieves higher performance with fewer parameters and comparable or lower FLOPs.

Qualitative results are provided in Fig. 2, showing superior segmentation performance across all tasks. For vessel segmentation in fundus images, our method accurately captures nearly all vessels, whereas other methods struggle to recover complex tree like structures and often miss many branches. In polyp, skin, thyroid, and cell segmentation, our approach better preserves region shapes while avoiding over segmentation, a common issue observed in several CNN and Mamba-based methods. For multi-class segmentation on the ACDC and Synapse datasets, most methods produce reasonable results; however, CNN based models often miss regions, while EMCAD and MCADS occasionally fail to detect certain classes.

5 Ablation Studies

In this section, we conduct ablation studies to analyze the key architectural components and design choices of the proposed decoder, isolating their individual and combined contributions to segmentation performance through systematic empirical evaluation. All experiments are performed on the Synapse multi-organ dataset for multi-class segmentation and the CellSeg dataset for binary segmentation to ensure reliable evaluation across settings.

5.1 Component-Level Analysis of Decoder Architectural Design Choices

We conduct a component-level ablation to analyze the individual and cumulative contributions of each decoder module, as summarized in Table 4. The baseline employs a PVTv2-B2 encoder with a plain U-Net style decoder, revealing the limitations of naive upsampling and direct skip fusion. Adding BT improves performance by injecting global context at the deepest stage. Enabling CA without RA further enhances accuracy through multi-scale aggregation, though gains remain limited due to the lack of explicit global modulation. Incorporating RA within CA yields a larger improvement, highlighting the importance of residual global stabilization. Adding RB strengthens reconstruction, and integrating SCA-Gate delivers a clear boost, showing that competitive and structured skip regulation is more effective than direct concatenation. Finally, we

evaluate Stage 0, which introduces an additional refinement pathway from the raw input. It applies RB and SCA-Gate after the Stage 1 2x upsampling. Although it yields slight improvements, the gains are marginal and inconsistent across tasks, and the total computational cost increases to 8.317 GFLOPs. Given this unfavorable trade off, Stage 0 is not included in the final model. Fig. 3 shows a radar plot of Deep Supervision (DS) and Edge Supervision (ES) settings across six benchmarks. In the Dice plot, enabling both DS and ES covers the largest area, showing the best segmentation accuracy. In the HD95 plot, the same setting covers the smallest area, indicating lower boundary error and better contour accuracy. Overall, performance improves consistently with BT, CA with RA, RB, and SCA-Gate. Accordingly, the final decoder configuration is directly guided by the empirical evidence, where each retained component demonstrates consistent and complementary gains.

5.2 Comparison with Baseline Attention Mechanisms for Skip Connection

To validate SCA-Gate, we compare it with representative skip attention mechanisms in Table 5, where the baseline is our full model without any attention gate. The baseline already performs strongly, confirming that gains are not solely due to the backbone as evident from Table 4. Attention U-Net Gate yields a moderate improvement but remains limited by its simple additive gating formulation. Attention U-Net Gate with ECA provides a slight additional gain through enhanced channel sensitivity, yet still lacks explicit multi-scale spatial competition.

Attention U-Net Gate with our ECA-MSP provides additional gain, though it remains within a conventional gating framework. LGAG from EMCAD offers competitive performance by enlarging local context through grouped convolution based gating, but primarily emphasizes spatial refinement. RLAB from MCADS Decoder delivers stable gains via residual based refinement, yet does not explicitly model competitive encoder-decoder alignment. In contrast, SCA-Gate achieves the highest overall performance by jointly modeling spatial competition and channel aware contextual modulation for selective and semantically aligned skip transmission.

5.3 Backbone Variants and Resolution Analysis

Table 6 compares pretrained PyTorch timm encoders under different input resolutions, where higher resolution improves segmentation accuracy but significantly increases

Table 3: Performance comparison with SOTA methods on the ACDC dataset. Overall Dice, IoU, and HD95 are reported together with per class Dice scores. All methods were reproduced and averaged over five runs, with fine tuning applied to SOTA models for fair comparison.

Method	Dice $\uparrow$	IoU $\uparrow$	HD95 $\downarrow$	RV	Myo	LV
U-Net [40]	81.56	73.41	6.9854	76.99	80.28	87.43
AttnUNet [33]	82.37	73.94	6.1684	78.13	81.08	87.89
UNet++ [66]	81.97	73.92	6.4724	77.74	80.73	87.44
nnU-Net [20]	82.66	74.27	6.1663	79.00	81.01	87.97
PraNetV2 [16]	83.74	76.13	6.3719	79.61	83.10	88.51
TransUNet [6]	83.07	74.85	5.7578	79.16	81.65	88.41
Swin-Unet [4]	82.61	74.59	6.1244	78.94	80.17	88.73
UCTransNet [49]	84.89	77.57	5.6995	80.94	84.11	89.62
U-Mamba [29]	84.18	76.47	5.8501	80.90	83.24	88.40
VM-UNet [41]	81.02	72.74	7.0025	76.75	79.40	86.90
EMCAD [37]	85.07	77.73	5.2472	81.58	84.23	89.42
MCADS [57]	84.51	76.92	5.5595	81.16	83.27	89.09
Ours	87.54 $\pm$ 0.3	80.96	4.4057	85.27	86.23	91.11

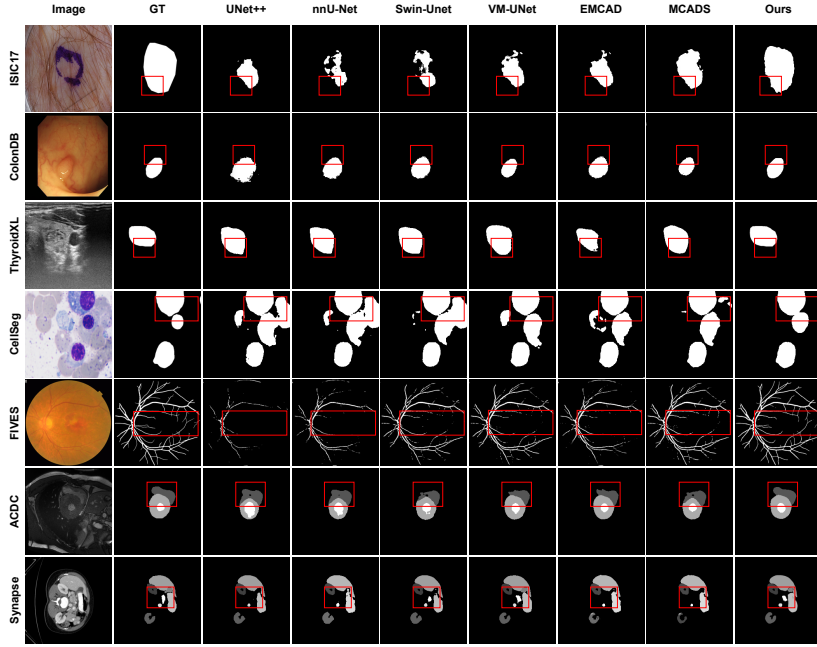


Fig. 2: Qualitative Results Comparison. Red rectangles highlight incorrect segmentation regions.

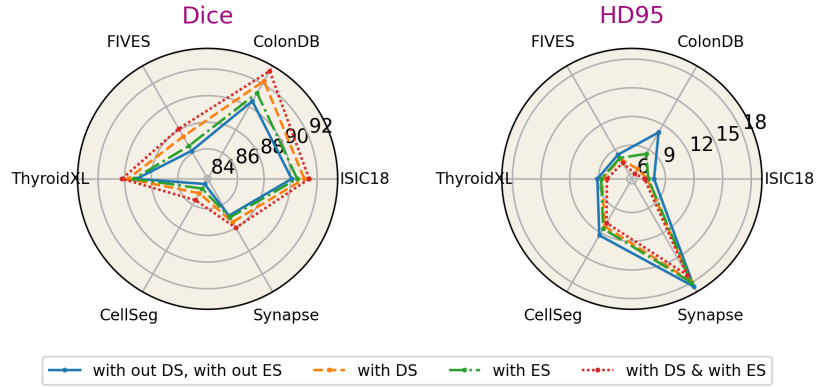


Fig. 3: Radar plots showing the effect of Deep Supervision (DS) and Edge Supervision (ES) across six segmentation benchmarks. In the Dice plot ($\uparrow$), performance improves as values move toward the outer rings. In the HD95 plot ($\downarrow$), lower values are better, so profiles closer to the center indicate more accurate boundaries. Compared with using either supervision alone or neither, enabling both DS and ES consistently achieves the best overall performance across all datasets.

computational cost, As expected, performance scales with encoder strength while remaining competitive even with lighter backbones, underscoring that our claim is compatibility rather than identical performance across encoder families. Convnext shows lower accuracy, maxvit achieves higher accuracy, and swin again attains relatively lower performance, while all remain heavier than the selected backbone. At 512x512 resolution, pvt_v2 improves accuracy over most 224 settings, and maxvit achieves the highest performance overall, but both require substantially greater computational resources. In comparison, our selected pvt_v2_b2 with 224x224 input provides a more practical balance between performance and efficiency, as higher resolution variants offer gains at the cost of substantial computational overhead.

Table 4: Ablation study of decoder components. Average Dice scores $\uparrow$ are reported.

Bottleneck	Context Aggregator w RA	Context Aggregator w/o RA	Refinement Block	SCA-Gate	Stage 0	Synapse	CellSeg
$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	73.91	81.07
$\checkmark$	$\times$	$\times$	$\times$	$\times$	$\times$	75.53	82.80
$\checkmark$	$\times$	$\checkmark$	$\times$	$\times$	$\times$	79.38	82.60
$\times$	$\checkmark$	$\times$	$\times$	$\times$	$\times$	81.03	82.47
$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	$\times$	83.57	84.28
$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\times$	$\times$	85.19	84.62
$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	<u>87.00</u>	<u>86.61</u>
$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	87.21	86.63

Table 5: Comparative Analysis of Skip Attention Modules. Average Dice scores $\uparrow$ are reported.

Skip Connection Attention	Params	Flops	Synapse	CellSeg
Attention U-Net - Gate	29.54 M	4.626 G	85.20	83.19
Attention U-Net - Gate + ECA	29.58 M	4.625 G	85.38	80.77
Attention U-Net - Gate + ECA-MSP	29.45 M	4.626 G	<u>85.46</u>	<u>84.65</u>
LGAG (EMCAD)	30.94 M	4.991 G	84.51	81.99
RLAB (MCADS-Decoder)	30.96 M	5.861 G	85.17	80.37
SCA-Gate (Ours)	30.60 M	5.001 G	87.00	86.61

Table 6: Comparison of Different Encoder Backbones. Average Dice scores $\uparrow$ are reported.

Encoder	Input Size	Params	Flops	Synapse	CellSeg
convnext_base	224x224	93.72 M	16.538 G	86.79	84.13
maxvit_base_tf_224	224x224	<u>84.19</u> M	50.317 G	87.57	<u>86.79</u>
swin_base_patch4_window7_224	224x224	92.89 M	<u>16.622</u> G	86.52	85.50
pvt_v2_b2	512x512	30.61 M	67.495 G	87.99	86.75
maxvit_base_tf_512	512x512	84.57 M	179.784 G	88.95	87.94

6 Conclusion

In this work, we revisit medical image segmentation from a decoder-centric perspective. We introduced a context-aware gated decoding framework for medical image segmentation that integrates global context modeling and adaptive skip fusion within a unified encoder-decoder design. By refining multi-scale features in a shared space, the method improves semantic consistency and boundary accuracy while maintaining strong computational efficiency compared to recent SOTA approaches, achieving consistent gains across diverse benchmarks. While this work focuses on improving decoder accuracy and efficiency under standardized 2D evaluation protocols, extending the framework to OOD robustness, 3D segmentation, and comparisons with large scale foundation models remains an important direction for future research.

7 Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00573160), the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2026-RS-2020-II201489), and the “Advanced GPU Utilization Support Program” funded by the Government of the Republic of Korea (Ministry of Science and ICT).

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. *Data in brief* **28**, 104863 (2020). <https://doi.org/10.1016/j.dib.2019.104863>
2. Bernard, O., Lalonde, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging* **37**(11), 2514–2525 (2018). <https://doi.org/10.1109/TMI.2018.2837502>
3. Bui, N.T., Hoang, D.H., Tran, M.T., Doretto, G., Adjeroh, D., Patel, B., Choudhary, A., Le, N.: Sam3d: Segment anything model in volumetric medical images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2024). <https://doi.org/10.48550/arXiv.2309.03493>
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022). <https://doi.org/10.48550/arXiv.2105.05537>
5. Chen, G., Liu, Y., Qian, J., Zhang, J., Yin, X., Cui, L., Dai, Y.: Dseu-net: A novel deep supervision seu-net for medical ultrasound image segmentation. *Expert Systems with Applications* **223**, 119939 (2023). <https://doi.org/10.1016/j.eswa.2023.119939>
6. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M.P., Zhang, S., Xing, L., Lu, L., Yuille, A., Zhou, Y.: Transunet: Rethinking the

- u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis* **97**, 103280 (2024). <https://doi.org/10.48550/arXiv.2102.04306>
7. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 801–818 (2018). <https://doi.org/10.48550/arXiv.1802.02611>
 8. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368* (2019). <https://doi.org/10.48550/arXiv.1902.03368>
 9. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. pp. 168–172. IEEE (2018). <https://doi.org/10.1109/ISBI.2018.8363547>
 10. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research* **2** (2023). <https://doi.org/10.26599/AIR.2023.9150015>
 11. Duong, V.H., Vu, H., Phan, H.D., Nguyen, D.Q., Pham, D.H., Le, Q.T., Nguyen, B.S., Do, T.D., Dinh, V.S., Nguyen, T.C., et al.: Thyroidxl: Advancing thyroid nodule diagnosis with an expert-labeled, pathology-validated dataset. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 616–626. Springer (2025). https://doi.org/10.1007/978-3-032-05182-0_60
 12. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: *International conference on medical image computing and computer-assisted intervention*. pp. 263–273. Springer (2020). <https://doi.org/10.48550/arXiv.2006.11392>
 13. Gao, Y., Jiang, Y., Peng, Y., Yuan, F., Zhang, X., Wang, J.: Medical image segmentation: A comprehensive review of deep learning-based methods. *Tomography* **11**(5), 52 (2025). <https://doi.org/10.3390/tomography11050052>
 14. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: *First conference on language modeling* (2024). <https://doi.org/10.48550/arXiv.2312.00752>
 15. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 574–584 (2022). <https://doi.org/10.48550/arXiv.2103.10504>
 16. Hu, B.C., Ji, G.P., Shao, D., Fan, D.P.: Pranet-v2: Dual-supervised reverse attention for medical image segmentation. *Computational Visual Media* (2026). <https://doi.org/10.48550/arXiv.2504.10986>
 17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7132–7141 (2018). <https://doi.org/10.48550/arXiv.1709.01507>
 18. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. pp. 1055–1059. Ieee (2020). <https://doi.org/10.48550/arXiv.2004.08790>

19. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging* **42**(5), 1484–1494 (2023). <https://doi.org/10.48550/arXiv.2109.07162>
20. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**(2), 203–211 (2021). <https://doi.org/10.48550/arXiv.1809.10486>
21. Jadon, S.: A survey of loss functions for semantic segmentation. In: 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB). pp. 1–7. IEEE (2020). <https://doi.org/10.1109/CIBCB48159.2020.9277638>
22. Jiang, C., Ding, T., Song, C., Tu, J., Yan, Z., Shao, Y., Wang, Z., Shang, Y., Han, T., Tian, Y.: Medical sam3: A foundation model for universal prompt-driven medical image segmentation. *arXiv preprint arXiv:2601.10880* (2026). <https://doi.org/10.48550/arXiv.2601.10880>
23. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific data* **9**(1), 475 (2022). <https://doi.org/10.1038/s41597-022-01564-3>
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023). <https://doi.org/10.48550/arXiv.2511.16719>
25. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015). <https://doi.org/10.7303/syn3193805>
26. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., et al.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International conference on medical image computing and computer-assisted intervention. pp. 615–625. Springer (2024). <https://doi.org/10.48550/arXiv.2402.03302>
27. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022). <https://doi.org/10.48550/arXiv.2201.03545>
28. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nature Communications* **15**(1), 654 (2024). <https://doi.org/10.1038/s41467-024-44824-z>
29. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722* (2024). <https://doi.org/10.48550/arXiv.2401.04722>
30. Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., Lou, W., Li, H., Upschulte, E., Dickscheid, T., de Almeida, J.G., Wang, Y., Han, L., Yang, X., Labagnara, M., Gligorovski, V., Scheder, M., Rahi, S.J., Kempster, C., Pollitt, A., Espinosa, L., Mignot, T., Middeke, J.M., Eckardt, J.N., Li, W., Li, Z., Cai, X., Bai, B., Greenwald, N.F., Valen, D.V., Weisbart, E., Cimini, B.A., Cheung, T., Brück, O., Bader, G.D., Wang, B.: The multi-modality cell segmentation challenge: Towards universal solutions. *Nature Methods* **21**, 1103–1113 (2024). <https://doi.org/10.1038/s41592-024-02233-6>
31. Müller, D., Hartmann, D., Meyer, P., Auer, F., Rey, I.S., Kramer, F.: Miseval: A metric library for medical image segmentation evaluation. In: MIE. pp. 33–37 (2022). <https://doi.org/10.48550/arXiv.2201.09395>
32. Ni, Z.L., Bian, G.B., Zhou, X.H., Hou, Z.G., Xie, X.L., Wang, C., Zhou, Y.J., Li, R.Q., Li, Z.: Raunet: Residual attention u-net for semantic segmentation of cataract surgical in-

- struments. In: International Conference on Neural Information Processing. pp. 139–149. Springer (2019). <https://doi.org/10.48550/arXiv.1909.10360>
33. Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. In: Medical Imaging with Deep Learning (2018). <https://doi.org/10.48550/arXiv.1804.03999>
34. Pu, Q., Xi, Z., Yin, S., Zhao, Z., Zhao, L.: Advantages of transformer and its application for medical image segmentation: a survey. *BioMedical engineering online* **23**(1), 14 (2024). <https://doi.org/10.1186/s12938-024-01212-4>
35. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition* **106**, 107404 (2020). <https://doi.org/10.48550/arXiv.2005.09007>
36. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6222–6231 (January 2023). <https://doi.org/10.1109/WACV56688.2023.00616>
37. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11769–11779 (2024). <https://doi.org/10.48550/arXiv.2405.06880>
38. Rahman, M.M., Shokouhmand, S., Bhatt, S., Faezipour, M.: Mist: Medical image segmentation transformer with convolutional attention mixing (cam) decoder. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 404–413 (2024). <https://doi.org/10.48550/arXiv.2310.19898>
39. Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked* **47**, 101504 (2024). <https://doi.org/10.1016/j.imu.2024.101504>
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015). <https://doi.org/10.48550/arXiv.1505.04597>
41. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024). <https://doi.org/10.48550/arXiv.2402.02491>
42. Shaharabany, T., Dahan, A., Giryas, R., Wolf, L.: Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370* (2023). <https://doi.org/10.48550/arXiv.2306.06370>
43. Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging* **23**(4), 501–509 (2004). <https://doi.org/10.1109/TMI.2004.825627>
44. Titoriya, A., Singh, M.: Pvt-cascade network on skin cancer dataset. In: 8th International Conference on Computing in Engineering and Technology (ICCET 2023). vol. 2023, pp. 480–486. IET (2023). <https://doi.org/10.1049/icp.2023.1536>
45. Tomar, N.K., Jha, D., Ali, S., Johansen, H.D., Johansen, D., Riegler, M.A., Halvorsen, P.: Ddanet: Dual decoder attention network for automatic polyp segmentation. In: International conference on pattern recognition. pp. 307–314. Springer (2021). <https://doi.org/10.48550/arXiv.2012.15245>
46. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: European conference on computer vision. pp. 459–479. Springer (2022). <https://doi.org/10.48550/arXiv.2204.01697>

47. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: International conference on medical image computing and computer-assisted intervention. pp. 23–33. Springer (2022). <https://doi.org/10.48550/arXiv.2203.04967>
48. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering* **2017**(1), 4037190 (2017). <https://doi.org/10.1155/2017/4037190>
49. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2441–2449 (2022). <https://doi.org/10.48550/arXiv.2109.04335>
50. Wang, H., Cao, P., Yang, J., Zaiane, O.: Narrowing the semantic gaps in u-net with learnable skip connections: The case of medical image segmentation. *Neural Networks* **178**, 106546 (2024). <https://doi.org/10.48550/arXiv.2312.15182>
51. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.: Understanding convolution for semantic segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 1451–1460. Ieee (2018). <https://doi.org/10.48550/arXiv.1702.08502>
52. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020). <https://doi.org/10.48550/arXiv.1910.03151>
53. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021). <https://doi.org/10.48550/arXiv.2102.12122>
54. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media* **8**(3), 415–424 (2022). <https://doi.org/10.1007/s41095-022-0274-8>
55. Wazir, S., Fraz, M.M.: Histoseg: Quick attention with multi-loss function for multi-structure segmentation in digital histology images. In: 2022 12th International Conference on Pattern Recognition Systems (ICPRS). pp. 1–7. IEEE (2022). <https://doi.org/10.48550/arXiv.2209.00729>
56. Wazir, S., Kim, D.: Rethinking the nested u-net approach: Enhancing biomarker segmentation with attention mechanisms and multiscale feature fusion. In: International Conference on Medical Imaging and Computer-Aided Diagnosis. pp. 175–186. Springer (2024). https://doi.org/10.1007/978-981-96-3863-5_17
57. Wazir, S., Kim, D.: Rethinking decoder design: Improving biomarker segmentation using depth-to-space restoration and residual linear attention. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30861–30871 (2025). <https://doi.org/10.48550/arXiv.2506.18335>
58. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16133–16142 (2023). <https://doi.org/10.48550/arXiv.2301.00808>
59. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018). <https://doi.org/10.48550/arXiv.1807.06521>

60. Xie, B., Tang, H., Cai, D., Yan, Y., Agam, G.: Self-prompt sam: Medical image segmentation via automatic prompt sam adaptation. *arXiv preprint arXiv:2502.00630* (2025). <https://doi.org/10.48550/arXiv.2502.00630>
61. Xie, Y., Yang, B., Guan, Q., Zhang, J., Wu, Q., Xia, Y.: Attention mechanisms in medical image segmentation: A survey. *arXiv preprint arXiv:2305.17937* (2023). <https://doi.org/10.48550/arXiv.2305.17937>
62. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2403–2412 (2018). <https://doi.org/10.48550/arXiv.1707.06484>
63. Yuan, D., Xu, Z., Tian, B., Wang, H., Zhan, Y., Lukasiewicz, T.: mu-net: Medical image segmentation using efficient and effective deep supervision. *Computers in Biology and Medicine* **160**, 106963 (2023). <https://doi.org/10.1016/j.compbiomed.2023.106963>
64. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785* (2023). <https://doi.org/10.48550/arXiv.2304.13785>
65. Zhao, P., Zhang, J., Fang, W., Deng, S.: Scau-net: spatial-channel attention u-net for gland segmentation. *Frontiers in Bioengineering and Biotechnology* **8**, 670 (2020). <https://doi.org/10.3389/fbioe.2020.00670>
66. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: *International workshop on deep learning in medical image analysis*. pp. 3–11. Springer (2018). <https://doi.org/10.48550/arXiv.1807.10165>