Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prompt-Conditioned Channel Attention for Hierarchical Feature Modulation toward Anatomy-Agnostic Segmentation

Mosharof Hossain
Khulna University of Engineering and Technology
Md Rabiul Islam
Texas A&M University
Limon Halder
Khulna University of Engineering and Technology
Erchin Serpedin
Texas A&M University
Md Kamrul Hasan
k.hasan22@imperial.ac.uk

Imperial College London

Article

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9601914/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Prompt-Conditioned Channel Attention for Hierarchical Feature Modulation toward Anatomy-Agnostic Segmentation

Mosharof Hossain[1], Md Rabiul Islam[2], Limon Halder[1], Erchin Serpedin[2], and Md Kamrul Hasan[3,1,*]

[1]Department of Electrical and Electronic Engineering (EEE), Khulna University of Engineering & Technology (KUET), Khulna-9203, Bangladesh
[2]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA
[3]Department of Bioengineering, Imperial College London, London SW7 2AZ, UK
[*]Correspondence to: k.hasan22@imperial.ac.uk; kamruleeekuet@gmail.com

ABSTRACT

Accurate and anatomically plausible medical image segmentation remains challenging due to low contrast, ambiguous boundaries, inter-patient variability, and modality-specific artifacts. Interactive segmentation using spatial prompts has emerged as a promising strategy to guide attention and improve localization, particularly in low-contrast or structurally ambiguous regions. However, existing methods typically restrict prompt integration to late-stage fusion and lack explicit mechanisms for channel-wise feature modulation, limiting their ability to capture deeper contextual and modality-specific variations across hierarchical representations. To address these limitations, we introduce *Prompt-Conditioned Channel Attention* (PCCA), a novel modulation mechanism that enables deep, hierarchical integration of semantic prompts within encoder-decoder segmentation networks. PCCA extracts compact channel descriptors from both image and prompt features via pooling, projects them into a shared latent space, and fuses them through a gated excitation unit to compute prompt-aware channel attention weights. These weights perform hierarchical channel-wise attention by adaptively reweighting feature responses at every network stage, enabling spatially selective and semantically enriched representations. Building on this mechanism, we propose **PROMISE-Net** (**PRO**mpt **M**odulated **I**ntegration for **SE**gmentation), instantiated in two architectural variants: a convolutional model (**PROMISE-C$_{\mathbf{NN}}$**) and a transformer-based model (**PROMISE-T$_{\mathbf{xformer}}$**). Across ISIC-Lesion, Kvasir-Polyp, CAMUS-Cardiac, and Kvasir-Instrument benchmarks, integrating PCCA into PROMISE-C$_{\text{NN}}$ yielded relative IoU improvements of *10.4%*, *8.7%*, *0.8%*, and *3.4%* over baseline U-Net, while PROMISE-T$_{\text{xformer}}$ achieved corresponding gains of *7.6%*, *23%*, *2.1%*, and *1.1%* over baseline UNETR. These results demonstrate consistent cross-architectural, cross-modal, and cross-anatomical improvements, establishing PCCA coupled PROMISE-Net as a scalable and generalizable framework for prompt-aware feature modulation in medical image segmentation. The code is available at https://github.com/kamruleee51/PROMISENet.

1 Introduction

Medical image segmentation plays a crucial role in modern healthcare, offering significant benefits in clinical diagnosis, surgical planning, and treatment monitoring[1]. Accurately delineating anatomical structures and pathological regions enables more informed clinical decision-making and contributes to improved patient outcomes. Nevertheless, achieving reliable segmentation remains highly challenging due to the inherent complexity and variability of medical images. These challenges arise primarily from two major aspects. First, medical images exhibit considerable diversity in the shape, size, and spatial configuration of anatomical structures. For instance, in skin lesion analysis, imaging artifacts such as hair, blood vessels, and uneven illumination can obscure lesion boundaries and confound interpretation. In addition, irregular contours and heterogeneous intensity distributions further degrade the performance of conventional shape- or intensity-based models[2,3]. Second, image resolution exerts a notable influence on segmentation performance. Low-resolution scans often suffer from poor boundary definition and loss of fine anatomical detail, whereas excessively high-resolution images may amplify noise and introduce unstable feature responses.

Manual delineation by clinical experts, though highly accurate, is both labor-intensive and time-consuming, rendering it impractical for large-scale clinical deployment. To mitigate this, a wide range of automatic segmentation algorithms have been developed over the past decades[4,5]. Early approaches relied on classical image processing techniques such as thresholding[6] and edge detection[7]. However, these methods lacked robustness and generalizability when confronted with the high anatomical variability and complex appearance patterns present in real-world medical images. Recent advances in deep learning (DL), particularly the advent of convolutional neural networks (CNNs)[8] and Transformers[9], have revolutionized medical image

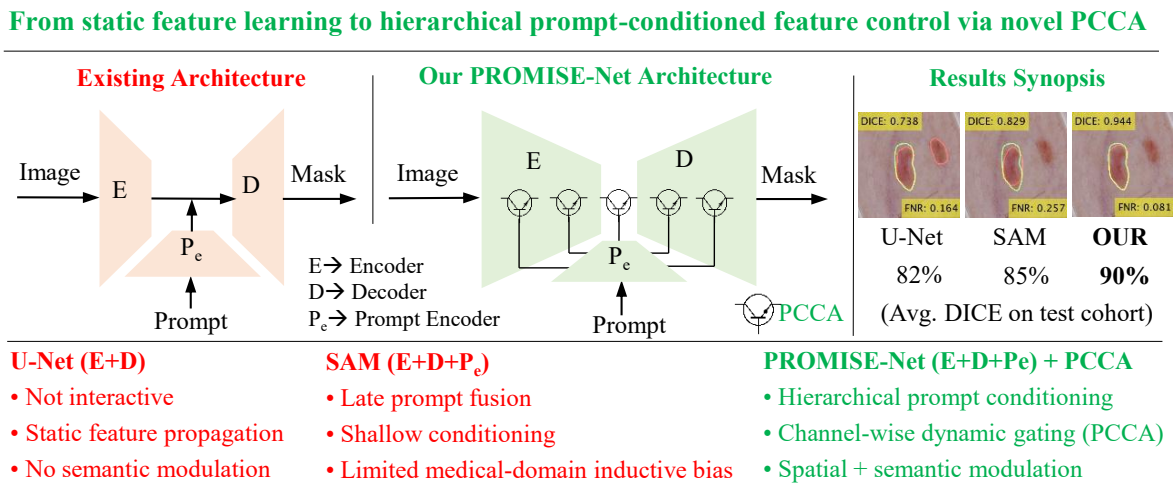


Figure 1. From static feature learning to hierarchical prompt-conditioned feature control. Compared with U-Net and SAM, which rely on static representations or shallow prompt fusion, PROMISE-Net integrates Prompt-Conditioned Channel Attention (PCCA) throughout the network, enabling dynamic channel-wise gating and spatial–semantic modulation for improved segmentation performance.

segmentation, achieving state-of-the-art results across diverse modalities and anatomical targets[10,11]. Despite their success, these models typically demand large-scale, fully annotated datasets to attain optimal accuracy and generalization. Acquiring such expert annotations is both costly and time-intensive, as it requires domain-specific knowledge and meticulous manual effort.

To alleviate the reliance on dense annotations, weakly and semi-supervised segmentation strategies have been widely explored, leveraging sparse or noisy supervisory signals such as image-level tags, bounding boxes, scribbles, or points[12–17]. Although these methods substantially reduce annotation costs, their limited spatial supervision often leads to inaccurate boundary delineation and structural inconsistencies, particularly in low-contrast or artifact-prone regions. To mitigate these shortcomings, several studies have incorporated structural priors, adversarial shape regularization, or consistency constraints to improve spatial coherence[18,19]. Yet, these techniques primarily enforce implicit shape regularity and fail to provide explicit spatial guidance that adapts to heterogeneous imaging conditions and anatomical variability.

Among the weakly supervised paradigms, scribble supervision has proven particularly effective due to its ease of use and practicality for delineating complex anatomical structures[20]. Annotators need only draw a few representative strokes per class, drastically reducing annotation time. However, the sparse and localized nature of scribble annotations provides limited supervision in low-contrast or ambiguous regions, often leading to segmentation errors in internal substructures and degraded boundary precision. Conventional architectures such as U-Net struggle to reconstruct missing contextual information, resulting in incomplete shape recovery of critical anatomical components. These challenges underscore the need for frameworks capable of learning effectively from sparse supervision while preserving detailed anatomical morphology.

To strengthen the supervision signal, pseudo-labeling (PL) has emerged as an effective enhancement within scribble-guided segmentation. PL methods exploit model predictions to generate provisional labels for unlabeled regions, thereby expanding the effective training set[21–23]. Typically, pseudo-labels are aggregated from multiple networks or decoder branches to improve robustness. For instance, Luo et al.[21] dynamically blend outputs from dual decoders to diversify pseudo-label generation, while Wu et al.[24] select high-confidence pixels from ensemble confidence maps. Similarly, Han et al.[25] employ multi-branch architectures with varied dilation rates and feature-level perturbations, and PacingPseudo[26] enforces prediction consistency through a twin-network design. In addition,[27] integrates class activation maps to inject class-specific cues derived from scribble annotations. Despite these advances, most PL frameworks rely on simple averaging or random weighting during aggregation, which can amplify uncertainty and introduce label noise[28]. As a result, pseudo-labels often lack structural coherence and fail to guide fine-grained boundary refinement effectively.

In parallel, interactive and prompt-based segmentation has gained momentum as a powerful means of injecting spatial priors directly into deep models. In this paradigm, users provide minimal yet informative cues such as points, bounding boxes, or scribbles to guide model attention and localization[29] (shown in Fig. 1). This strategy is particularly attractive for human-in-the-loop clinical workflows, where clinicians can iteratively refine results while preserving the efficiency of automated inference. Despite these advantages, most existing prompt-driven frameworks, including SAM[29], incorporate prompts only through

late-stage fusion or single-layer conditioning, confining their influence to shallow or bottleneck representations. Moreover, the mechanisms by which prompt features interact with mainstream image features, especially at the level of channel-wise modulation, are often ad hoc or underspecified. As a result, prompt semantics are not effectively propagated across the network hierarchy, limiting the model's ability to shape intermediate representations. This restriction hinders the capture of multi-scale contextual relationships, modality-specific artifacts, and fine anatomical structures, particularly in low-contrast or structurally ambiguous regions.

In this work, we propose the following key contributions. First, we introduce *Prompt-Conditioned Channel Attention (PCCA)*, a novel attention mechanism that enables deep, hierarchical modulation of spatial prompts within encoder–decoder segmentation networks, allowing semantic guidance to propagate consistently throughout the feature hierarchy (illustrated in Fig. 1). Unlike existing prompt-driven designs that condition only shallow or bottleneck layers, PCCA embeds prompts directly into the feature transformation process at every stage. PCCA is a lightweight, prompt-conditioned channel excitation mechanism that internally executes squeeze-and-excitation operations to adaptively fuse prompt-derived spatial priors with visual features, thereby enhancing region-of-interest focus, boundary localization, and anatomical consistency while preserving computational efficiency. Second, we integrate PCCA into a unified segmentation framework, **PROMISE-Net** (**PRO**mpt **M**odulated **I**ntegration for **SE**gmentation), instantiated in two complementary variants: a convolutional model (**PROMISE-C$_{\mathbf{NN}}$**) and a transformer-based model (**PROMISE-T$_{\mathbf{xformer}}$**). This design demonstrates the cross-architectural generality and plug-in flexibility of the proposed mechanism. Finally, extensive evaluations on the ISIC-Lesion, CAMUS-Cardiac, Kvasir-Polyp, and Kvasir-Instrument benchmarks show that PROMISE-Net consistently achieves cross-architectural, cross-modal, and cross-anatomical improvements over strong state-of-the-art baselines. Collectively, these contributions establish PCCA coupled PROMISE-Net as a scalable and generalizable framework for prompt-aware feature modulation in medical image segmentation.

2 Results and Discussionin

2.1 PROMISENet Design and Ablation Analysis

We perform ablation experiments to assess the impact of integrating the proposed PCCA module at different stages of PROMISE-T$_{\text{xformer}}$ on two datasets: ISIC-2017 and CAMUS-Cardiac. As summarized in Table 1, these experiments quantify both segmentation performance and computational complexity when PCCA is inserted at the bottleneck alone or jointly at the bottleneck and decoder stages. For comparison, we also implement an alternative fusion scheme, termed outer product fusion (OPF), which replaces the global average pooling and shared latent projection in PCCA with a simple element-wise multiplication between image and prompt embeddings.

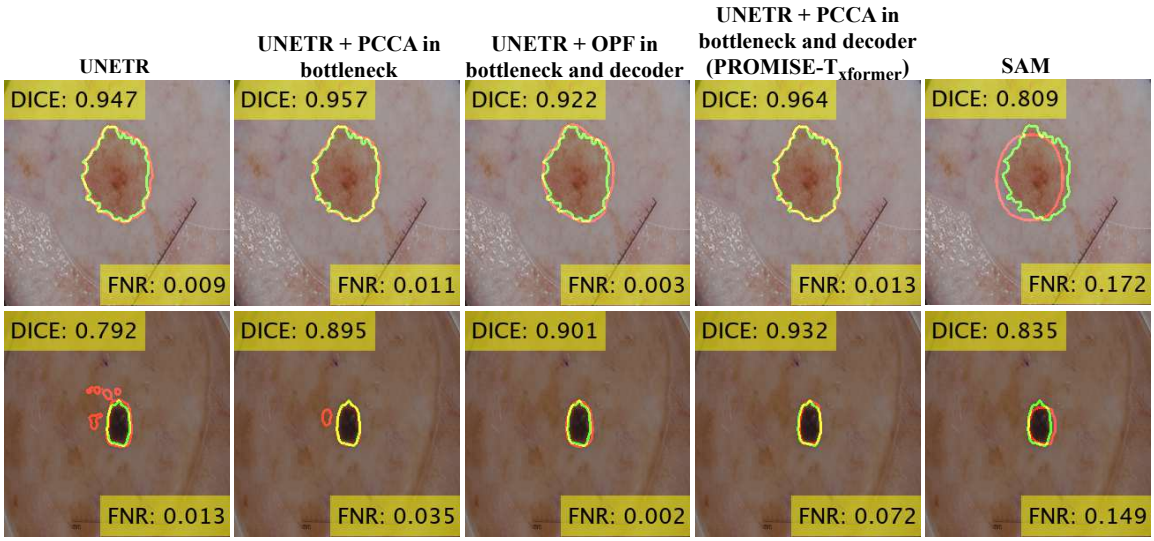
Table 1. Ablation study of PCCA integration on the ISIC-2017 and CAMUS datasets. Progressive insertion of PCCA into deeper UNETR stages improves DSC and reduces FNR with modest computational overhead, demonstrating that hierarchical prompt-guided modulation enhances feature discrimination and robustness. Bold blue values denote the best performance.

Methods	**Params**(↓)	**FLOPs**(↓)	**ISIC 2017**		**CAMUS**	
			DSC (%) (↑)	**FNR** (%) (↓)	**DSC** (%) (↑)	**FNR** (%) (↓)
(1) UNETR[9]	116,048,130	52,792,357,120	82.5 ± 18.0	19.0 ± 21.5	88.6 ± 5.4	11.6 ± 7.4
(2) UNETR + PCCA in bottleneck	127,922,994	58,231,075,072	88.3 ± 9.6	12.7 ± 13.1	89.8 ± 4.7	10.5 ± 6.2
(3) UNETR + PCCA in bottleneck and decoder	128,357,120	60,044,940,800	**89.0 ± 8.2**	10.6 ± 10.9	**90.0 ± 5.1**	**9.2 ± 5.8**
(4) UNETR + OPF in bottleneck and decoder	127,055,010	60,042,342,656	87.4 ± 9.4	**10.2 ± 11.4**	89.6 ± 5.0	9.8 ± 6.1
(1) vs. (2) ($\mathbf{p < 0.05}$?)			✓	✓	✓	✓
(2) vs. (3) ($\mathbf{p < 0.05}$?)			✓	✓	✓	✓
(3) vs. (4) ($\mathbf{p < 0.05}$?)			✓	✗	✓	✓

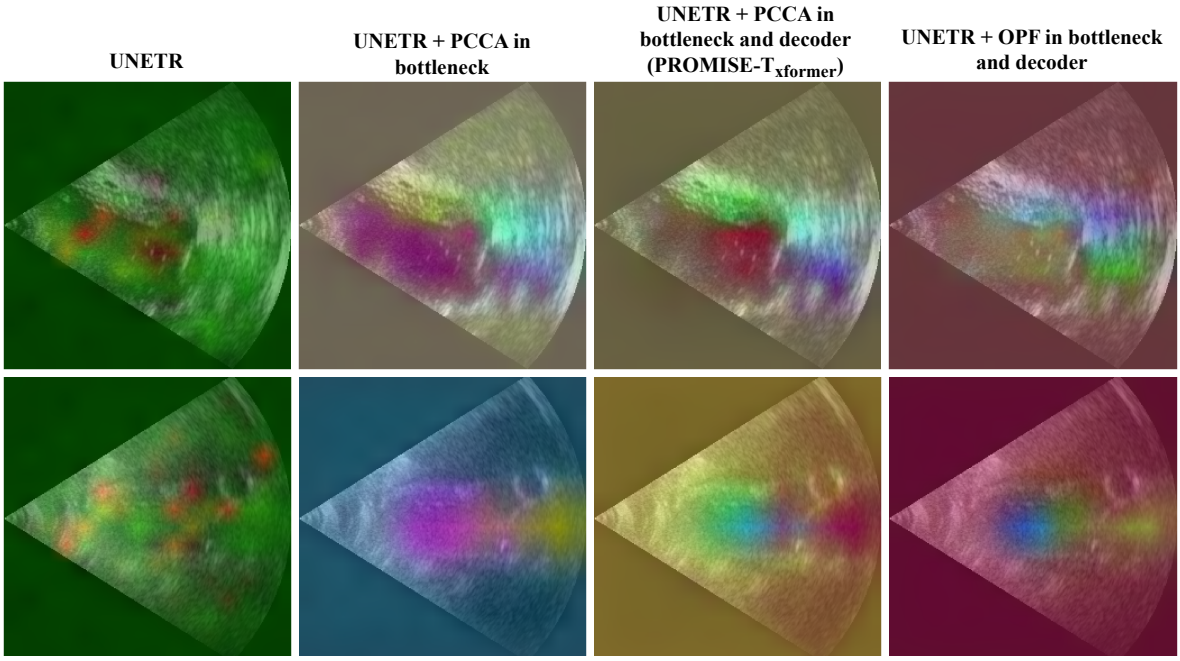
2.1.1 Impact of Hierarchical PCCA on Segmentation

Table 1 demonstrates that integrating PCCA into the UNETR backbone consistently improves segmentation accuracy across both ISIC 2017 and CAMUS datasets. Incorporating PCCA at the bottleneck stage yields a significant improvement on ISIC 2017, raising DSC from 82.5 to 88.3 (**+5.8%**, $p < 0.05$) and reducing FNR from 19.0 to 12.7 (**–6.3%**), while providing similar gains ($p < 0.05$) on CAMUS (**+2.8%** DSC and **–2.8%** FNR). Extending PCCA to both the bottleneck and decoder stages further increases segmentation fidelity ($p < 0.05$), achieving a DSC of **89.0%** and lowering the FNR to **10.6%** on ISIC 2017, while attaining **92.1%** DSC and **9.2%** FNR on CAMUS.

In contrast, the simple OPF module with UNETR improves the DSC and FNR for both the datasets compared with the baseline UNETR (Table 1), again confirming that introducing spatial prompts for feature modulation has a substantial impact on



(a) Qualitative results on ISIC 2017 (Green: ground truth and Red: prediction).



(b) Intermediate feature-map comparison for UNETR variants, visualized by PCA projection to RGB space.

Figure 2. Ablation visualization in two-column layout: PROMISE-Net with hierarchical PCCA yields more complete delineations and more spatially focused feature representations than baseline UNETR and OPF fusion.

segmentation performance. However, when compared with the proposed PCCA, the OPF variant remains significantly inferior overall, as PCCA achieves higher overlap accuracy and stronger cross-dataset generalization. For ISIC 2017, the reduction in FNR achieved by OPF is not statistically significant relative to that of PCCA, indicating that while OPF captures some prompt-conditioned information, it lacks the robust hierarchical modulation achieved by PCCA. The parameter and FLOP differences between variants remain minimal ($< 2\%$), confirming that the observed performance gains are primarily due to improved feature modulation by our novel PCCA rather than network scaling.

Visual comparisons in Fig. 2a are consistent with the quantitative results in Table 1 showing PROMISE-T$_{xformer}$'s better outputs with fewer discontinuities and reduced over-segmentation compared with the other baselines. Differences between methods are most apparent along lesion boundaries, where PROMISE-T$_{xformer}$ predictions more closely follow the ground

truth contours. Fig. 2a also includes results obtained using SAM. As reported in Table 3, SAM improves segmentation performance relative to the baseline UNETR; however, PROMISE-$T_{xformer}$ achieves higher overlap metrics than SAM on the evaluated datasets. These visual and quantitative comparisons indicate that the hierarchical integration of PCCA achieves the most balanced performance, enhancing segmentation robustness and boundary consistency with only marginal computational overhead.

2.1.2 Hierarchical Modulation and Feature Separation

Fig. 2b provides qualitative insight into Table 1 on the CAMUS dataset, visualizing intermediate feature maps for UNETR and its PCCA variants. The baseline UNETR exhibits diffuse feature responses with substantial overlap between color components, indicating limited separation between anatomical regions. Introducing PCCA at the bottleneck results in more spatially concentrated feature activations and improved separation of color regions. When PCCA is applied at both the bottleneck and decoder stages (PROMISE-$T_{xformer}$), corresponding to the highest quantitative performance on CAMUS (DSC 90.0%, FNR 9.2%), the feature embeddings appear more localized and exhibit clearer separation across distinct color components (Fig. 2b). In contrast, the OPF-based variant shows less consistent separation of feature responses compared with the hierarchical PCCA configuration. These visual patterns align with the quantitative trends in Table 1, where hierarchical PCCA integration yields higher overlap accuracy and lower false-negative rates than OPF.

2.1.3 Robustness to Channel Compression in PCCA

The reduction ratio r in PCCA controls the level of channel compression in the excitation MLP. Table 2 reports segmentation performance and model complexity on the ISIC 2017 dataset for $r \in \{2,4,8,16,32\}$. Across all configurations, HD95 varies within a narrow range (12.7–12.9), and FNR remains between 10.3% and 11.2%, indicating limited sensitivity to the choice of r. Among the evaluated settings, $r = 16$ yields the lowest FNR (10.3%) and a low HD95 (12.8) while maintaining moderate computational cost (52.15M parameters and 106.33G FLOPs). Parameter count and FLOPs decrease monotonically with increasing r, with differences remaining small across configurations. Based on these, $r = 16$ is used in all experiments.

Table 2. Effect of the reduction ratio r on PCCA efficiency and performance on ISIC-2017. A moderate setting (r=16) offers the best accuracy–efficiency trade-off, achieving low HD95 and FNR with minimal overhead. Bold blue indicates the best result.

Reduction Ratio (r)	**HD95** (↓)	**FNR** (%) (↓)	**FLOPs** (↓)	**Params** (↓)
$r = 2$	12.9	10.5	106.33347	53,070,338
$r = 4$	12.9	11.2	106.33242	52,545,794
$r = 8$	12.9	10.9	106.33190	52,283,522
$r = 16$	**12.8**	**10.3**	106.33164	52,152,386
$r = 32$	12.7	10.6	106.33150	52,086,818

2.2 Generalizability Analysis

2.2.1 Cross-Architectural Generalization

Cross-architectural generalization is evaluated by PROMISE-C_{NN} and PROMISE-$T_{xformer}$ backbones and comparing each against its corresponding baseline U-Net and UNETR across four datasets (Table 3).

For convolutional networks, PROMISE-C_{NN} improves DSC over U-Net by **+8.5%** on ISIC 2017 (82.2% to 90.7%), **+6.5%** on Kvasir-Polyp (86.0% to 92.5%), **+2.8%** on Kvasir-Instrument (94.4% to 97.2%), and **+0.5%** on CAMUS (90.8% to 91.3%). These gains are accompanied by substantial reductions in boundary error, including approximately **50%** on ISIC 2017 (23.6 to 11.7 pixels) and **55%** on Kvasir-Polyp (29.1 to 13.3 pixels), as well as consistent decreases in false-negative rate across all datasets. Qualitative examples in Fig. 3 (a,c) corroborate these trends, showing more complete delineations with reduced fragmentation relative to U-Net.

Similarly, PROMISE-$T_{xformer}$ consistently outperforms UNETR across all evaluated datasets. On ISIC 2017 and CAMUS, DSC increases from 82.5% to 89.0% and from 88.6% to 90.0%, respectively, with corresponding reductions in FNR. On Kvasir-Polyp and Kvasir-Instrument, PROMISE-$T_{xformer}$ achieves higher DSC and lower HD95 than UNETR, with statistically significant differences reported for most metrics. Visual comparisons in Fig. 3 (b,d) further show smoother contours and more continuous anatomical structures.

Together, the consistent quantitative gains across both convolutional and transformer-based backbones, supported by corresponding qualitative evidence, indicate that the performance gains introduced by PCCA are not specific to a particular network type.

Table 3. Performance comparison across four segmentation datasets: ISIC-2017 (skin lesion), Kvasir-Polyp, Kvasir-Instrument (endoscopic tool), and CAMUS (cardiac echocardiography). PROMISE-Net consistently outperforms baseline architectures across diverse anatomical structures and imaging modalities, demonstrating strong cross-domain and cross-architectural generalizability. Bold blue values indicate the best performance for each metric.

Methods	**DSC** (%) (↑)	**IoU** (%) (↑)	**HD95 (Pix)** (↓)	**FNR** (%) (↓)
Skin Lesion Dataset (ISIC-2017)[30]				
(1) U-Net[8]	82.2 ± 18.8	73.3 ± 21.8	23.6 ± 26.4	19.4 ± 21.9
(2) PROMISE-C_{NN} (proposed)	90.7 ± 7.5	83.7 ± 10.0	11.7 ± 10.6	7.3 ± 9.0
(1) vs. (2) ($\mathbf{p < 0.05}$?)	✓	✓	✓	✓
(3) UNETR[9]	82.5 ± 18.0	73.4 ± 21.4	22.3 ± 25.2	19.0 ± 21.5
(4) PROMISE-$T_{xformer}$ (proposed)	89.0 ± 8.2	81.0 ± 11.4	15.7 ± 18.2	10.6 ± 10.9
(3) vs. (4) ($\mathbf{p < 0.05}$?)	✓	✓	✓	✓
(5) SAM[29]	85.1 ± 9.0	75.0 ± 11.5	16.7 ± 10.7	12.1 ± 9.1
(2) vs. (5) ($\mathbf{p < 0.05}$?)	✓	✓	✓	✓
(4) vs. (5) ($\mathbf{p < 0.05}$?)	✓	✓	✓	✓
Polyp Segmentation Dataset (Kvasir-Polyp)[31]				
(1) U-Net[8]	86.0 ± 17.3	78.5 ± 20.8	29.1 ± 33.8	11.9 ± 19.7
(2) PROMISE-C_{NN} (proposed)	92.5 ± 9.8	87.2 ± 13.4	13.3 ± 18.6	7.3 ± 11.4
(1) vs. (2) ($\mathbf{p < 0.05}$?)	✓	✓	✓	✓
(3) UNETR[9]	68.3 ± 25.7	57.0 ± 27.0	58.0 ± 37.8	23.6 ± 27.7
(4) PROMISE-$T_{xformer}$ (proposed)	87.5 ± 14.8	80.0 ± 17.9	23.2 ± 33.9	9.0 ± 13.1
(3) vs. (4) ($\mathbf{p < 0.05}$?)	✓	✓	✓	✓
(5) SAM[29]	80.0 ± 13.1	68.2 ± 15.4	24.4 ± 24.0	14.7 ± 10.3
(2) vs. (5) ($\mathbf{p < 0.05}$?)	✓	✓	✓	✓
(4) vs. (5) ($\mathbf{p < 0.05}$?)	✓	✓	✗	✓
Diagnostic and Therapeutic Tool Segmentation (Kvasir-Instrument)[32]				
(1) U-Net[8]	94.4 ± 13.8	91.2 ± 14.8	8.7 ± 19.5	6.0 ± 15.3
(2) PROMISE-C_{NN} (proposed)	97.2 ± 1.4	94.6 ± 2.7	3.4 ± 3.0	2.7 ± 1.8
(1) vs. (2) ($\mathbf{p < 0.05}$?)	✗	✓	✓	✓
(3) UNETR[9]	95.3 ± 7.2	91.6 ± 9.8	8.1 ± 22.0	4.1 ± 6.9
(4) PROMISE-$T_{xformer}$ (proposed)	96.2 ± 2.2	92.7 ± 4.0	4.4 ± 2.5	4.5 ± 4.0
(3) vs. (4) ($\mathbf{p < 0.05}$?)	✗	✗	✗	✗
2D Echocardiography (CAMUS)[33]				
(1) U-Net[8]	90.8 ± 4.8	83.6 ± 7.3	6.2 ± 4.1	9.4 ± 6.6
(2) PROMISE-C_{NN} (proposed)	91.3 ± 4.1	84.4 ± 6.4	5.9 ± 4.1	8.4 ± 5.2
(1) vs. (2) ($\mathbf{p < 0.05}$?)	✓	✓	✗	✓
(3) UNETR[9]	88.6 ± 5.4	80.1 ± 8.1	7.9 ± 5.2	11.6 ± 7.4
(4) PROMISE-$T_{xformer}$ (proposed)	90.0 ± 5.1	82.2 ± 7.5	7.0 ± 5.2	9.2 ± 5.8
(3) vs. (4) ($\mathbf{p < 0.05}$?)	✓	✓	✓	✓

2.2.2 Cross-Anatomical and Cross-Domain Generalization

Cross-anatomical generalization is assessed across segmentation tasks involving distinct anatomical targets (Table 3), including skin lesions (ISIC 2017), gastrointestinal polyps (Kvasir-Polyp), surgical instruments (Kvasir-Instrument), and cardiac chambers

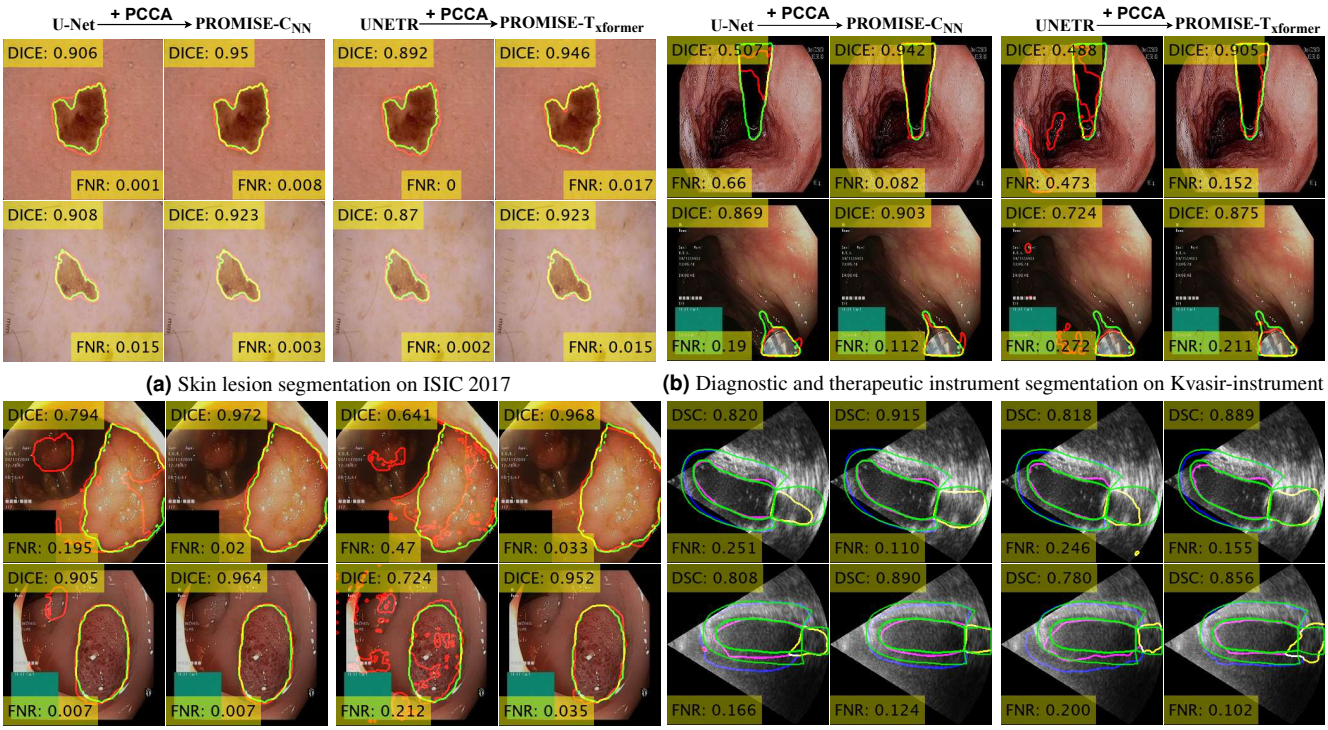


Figure 3. Comparison of baseline segmentation networks (U-Net and UNETR) and their PCCA-integrated counterparts, PROMISE-C$_{NN}$ and PROMISE-T$_{xformer}$, across cross-modal and cross-anatomical datasets. In (a)–(c), ground-truth boundaries are shown in green and predictions in red for binary segmentation tasks. In (d), for multi-class cardiac segmentation on CAMUS, green contours denote ground truth, while blue, purple, and yellow indicate predicted MYO, LV, and LA boundaries, respectively. PROMISE-Net variants exhibit improved anatomical fidelity, smoother contours, and reduced fragmentation compared to their baselines.

(CAMUS).

On ISIC 2017, PROMISE-C$_{NN}$ improves DSC from 82.2% to 90.7%, reduces HD95 from 23.6 to 11.7 pixels, and lowers FNR from 19.4% to 7.3%. On Kvasir-Polyp, DSC increases from 86.0% to 92.5%, with HD95 reduced from 29.1 to 13.3 pixels and FNR from 11.9% to 7.3%. For the anatomically distinct Kvasir-Instrument dataset, PROMISE-C$_{NN}$ improves DSC from 94.4% to 97.2%, reduces HD95 from 8.7 to 3.4 pixels, and lowers FNR from 6.0% to 2.7%. On the CAMUS dataset, DSC increases from 90.8% to 91.3%, with a corresponding reduction in FNR (Table 3). Qualitative results in Fig. 3 align with these quantitative improvements. PROMISE-Net variants produce lesion and polyp segmentations with fewer disconnected regions and closer adherence to annotated boundaries on ISIC 2017 and Kvasir-Polyp. In Kvasir-Instrument, PROMISE-Net more completely delineates elongated tool structures in regions affected by strong specular reflections, where baseline models exhibit partial detections or discontinuities. On CAMUS, PROMISE-Net yields smoother myocardial contours and clearer separation between cardiac chambers, consistent with the improved overlap accuracy and reduced false-negative rates reported in Table 3.

These results show that the prompt-conditioned modulation in PROMISE-Net transfers effectively across anatomically diverse targets without task-specific architectural modifications.

2.2.3 Cross Image Quality Levels and Cardiac Structures

Image quality variation robustness is evaluated on the CAMUS dataset by stratifying test samples into good, medium, and poor quality categories, as defined in the dataset annotations[33]. As shown in Fig. 5, PROMISE-T$_{xformer}$ achieves higher DSC than UNETR across all levels. The variation in DSC between good and poor quality images for PROMISE-T$_{xformer}$ remains within 5%, indicating limited sensitivity to image quality degradation relative to the baseline.

Class-wise analysis in Fig. 5 further demonstrates consistent improvements in DSC for all cardiac structures, including LV, MYO, and LA, with statistically significant differences reported for each class. Together with the binary segmentation results on ISIC-2017 and Kvasir datasets, these findings indicate that PROMISE-Net generalizes across label-space complexity, maintaining performance gains when transitioning from binary to multi-class segmentation tasks. Overall, the performance

Table 4. Quantitative comparison of the proposed PROMISE-C_{NN} with state-of-the-art segmentation methods across four datasets: ISIC-2017 (skin lesion), Kvasir-Polyp, Kvasir-Instrument (endoscopic tool), and CAMUS (2D echocardiography). Result on CAMUS is shown for the LA class. Bold blue values indicate the best performance for each metric.

(a) ISIC-2017 (Skin Lesion)

Methods	DSC (%) (↑)	IoU (%) (↑)	FNR (%) (↓)
Pact-Net[34]	86.2	79.3	13.8
USL-Net[35]	80.5	68.5	11.4
FAT-Net[36]	85.0	76.5	16.1
HTC-Net[37]	90.1	**84.0**	11.8
PROMISE-C_{NN}	**90.7**	83.7	**7.3**

(b) Kvasir-Polyp

Methods	DSC (%) (↑)	IoU (%) (↑)	FNR (%) (↓)
RM-UNet[38]	90.2	82.2	10.3
Pact-Net[34]	90.6	84.7	–
USL-Net[35]	91.2	85.9	–
ERDUNet[39]	90.7	84.6	–
PROMISE-C_{NN}	**92.5**	**87.2**	**7.3**

(c) Kvasir-Instrument

Methods	DSC (%) (↑)	IoU (%) (↑)	FNR (%) (↓)
RM-UNet[38]	94.7	89.9	5.9
MAF-Net[40]	96.8	**96.6**	–
DECA-Net[41]	96.9	93.9	–
ERDUNet[39]	95.2	91.6	–
PROMISE-C_{NN}	**97.2**	94.6	**2.7**

(d) CAMUS (2D Echocardiography)

Methods	DSC (%) (↑)	IoU (%) (↑)	HD95 (*mm*) (↓)
CLAS[42]	91.4	–	5
EchoSAM[43]	90.7	83.4	3.8
TAM-FCN8s[5]	91.6	–	**3.3**
CoST-UNet[44]	87.6	79.2	6.7
PROMISE-C_{NN}	**91.8**	**85.2**	3.9

Table 5. Computational efficiency, model complexity, and performance comparison of selected segmentation architectures. DSC is reported on ISIC-2017; bold blue denotes the best result per metric.

Model	**DSC**	**FLOPs**	**Params**
U-Net	82.2	96.6	31
SAM	85.1	2991.32	312
PROMISE-C_{NN}	90.7	118.15	61

gains of PROMISE-$T_{xformer}$ are preserved across both varying image quality conditions and multiple cardiac regions on the CAMUS dataset.

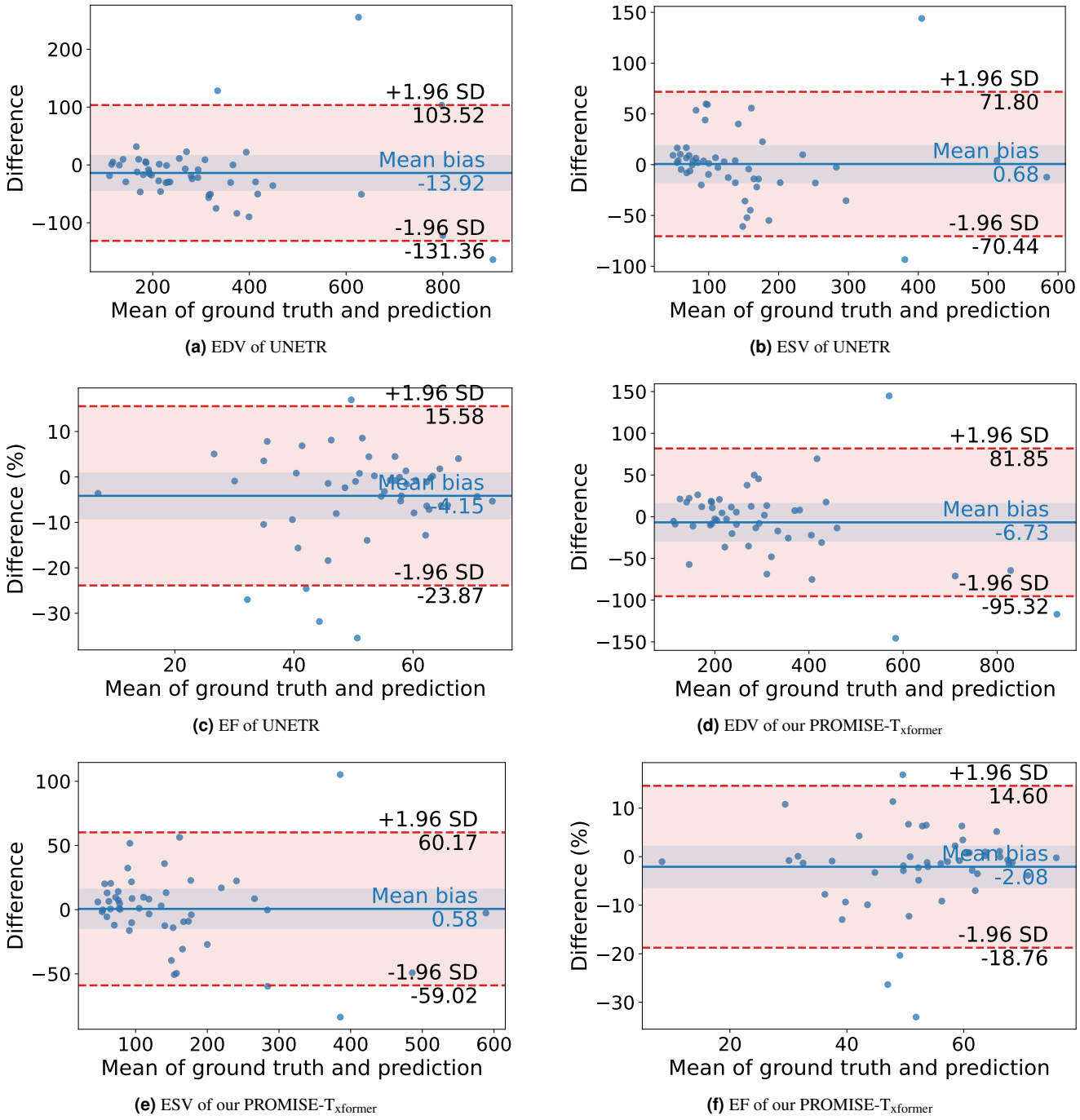
2.3 PCCA vs. Alternative SAM in Performance and Efficiency

PROMISE-Net is further compared with the alternative prompt-based SAM on the ISIC 2017 and Kvasir-Polyp datasets in Table 3. On ISIC-2017, PROMISE-C_{NN} achieves a DSC of 90.7% compared with 85.1% for SAM, while reducing HD95 from 16.7 to 11.7 pixels and FNR from 12.1% to 7.3%. On Kvasir-Polyp, PROMISE-C_{NN} attains a DSC of 92.5% versus 80.0% for SAM, with HD95 reduced from 24.4 to 13.3 pixels and FNR from 14.7% to 7.3%. PROMISE-$T_{xformer}$ similarly outperforms SAM in terms of DSC and FNR on both datasets, with statistically significant differences reported for most metrics (Table 3). These quantitative differences are consistent with the qualitative comparisons in Fig. 2a, where PROMISE-Net variants produce more complete and spatially coherent segmentations than SAM, indicating that hierarchical prompt-conditioned modulation provides advantages over late-stage prompt integration alone.

Again, as shown in Table 5, PROMISE-C_{NN} attains 90.7% DSC with 118.15 GFLOPs and 61 million parameters, outperforming U-Net (82.2%, 96.6 GFLOPs, 31 M) and remaining far lighter than SAM (85.1%, 2991.32 GFLOPs, 312 M). This balance of accuracy and computational cost makes PROMISE-Net suitable for practical deployment across diverse medical imaging modalities.

2.4 Robustness to User Prompt and Interaction Variability

To assess robustness to interaction variability, bounding-box prompts from two human observers were compared with automatically generated boxes on 300 ISIC-2017 test images. To simulate realistic user imprecision, ground-truth bounding boxes were synthetically perturbed by ± 20 pixels. PROMISE-Net exhibits comparable performance across observer-provided and automated prompts; pairwise effect sizes between observer-driven and automated prompts remain small (Cohen's $d < 0.2$), indicating negligible practical differences in the results (Fig. 6). These results demonstrate that PROMISE-Net is robust to inter-observer variability and moderate prompt perturbations, producing stable and reproducible segmentations despite differences in prompt placement or annotation style.



(a) EDV of UNETR

(b) ESV of UNETR

(c) EF of UNETR

(d) EDV of our PROMISE-$T_{xformer}$

(e) ESV of our PROMISE-$T_{xformer}$

(f) EF of our PROMISE-$T_{xformer}$

Figure 4. Bland–Altman plots comparing predicted and reference EDV, ESV, and EF for UNETR (a-c) and our PROMISE-$T_{xformer}$ (d-f). The proposed model reduced the mean bias (EDV: $-13.9 \rightarrow -6.7$ mL; EF: $-4.15 \rightarrow -2.08\%$) and narrowed the LoA across all indices, demonstrating improved volumetric precision and enhanced temporal coherence. The EF agreement (-18.8% to $+14.6\%$) lies within the inter-observer variability range, confirming clinically reliable ventricular function estimation.

2.5 Quantitative Agreement in Ventricular Volumes and Function

To assess quantitative agreement between estimated and reference measurements, Bland-Altman analyses were conducted for EDV, ESV, and EF on the CAMUS dataset (see Fig. 4). The baseline UNETR exhibited a systematic underestimation of EDV (mean bias $= -13.9$ mL, 95% LoA $= [-131.4, +103.5]$ mL) and wide LoA, indicating large inter-subject dispersion, particularly for subjects with dilated ventricles. The proposed PROMISE-$T_{xformer}$ reduced this bias to -6.7 mL and narrowed

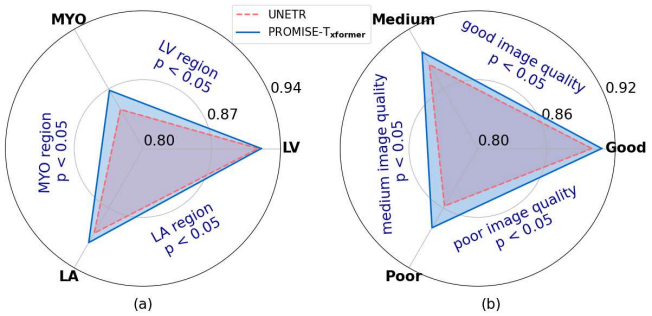


Figure 5. Comparison of segmentation performance between UNETR and PROMISE-T$_{\text{xformer}}$ on CAMUS: class-wise DSC (LV, MYO, and LA) and DSC across good, medium, and poor image quality levels[33].

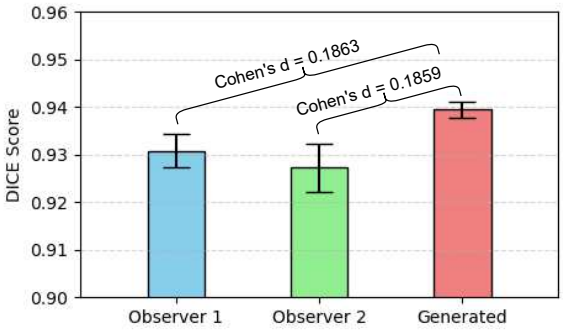


Figure 6. Dice score comparison using observer-drawn and automatically generated bounding-box prompts, with pairwise effect sizes (Cohen's d).

the LoA to $[-95.3, +81.9]$ mL, reflecting more accurate delineation of the endocardial surface at ED and improved volumetric consistency. For ESV, both models produced negligible bias (≈ 0.6 mL), but PROMISE-T$_{\text{xformer}}$ achieved tighter LoA (± 60 mL vs. ± 71 mL), suggesting higher precision during the contracted phase. The improvement was most evident for EF (improving the correlation from 72.9% to 80.2%), where the mean bias decreased from -4.15% to -2.08%, and the LoA contracted from $[-23.9\%, +15.6\%]$ to $[-18.8\%, +14.6\%]$. The smaller EF dispersion indicates enhanced temporal coherence between ED and ES segmentations, attributable to the phase-consistent contextual attention mechanism. No proportional bias was observed ($|r| < 0.1$, $p > 0.05$), confirming stable performance across the physiological range of ventricular sizes and functions. The EF limits fall within the reported inter-observer variability of expert annotations ($\sim \pm 15\%$), indicating that PROMISE-T$_{\text{xformer}}$ provides clinically interchangeable functional estimation. Collectively, these findings demonstrate that integrating interactive prompts through PCCA attention enhances not only segmentation fidelity but also the reproducibility of physiologically meaningful volumetric and functional indices.

2.6 Benchmark with State-of-the-Art

We benchmark PROMISE-Net against a diverse set of recent state-of-the-art methods, encompassing both convolutional and transformer-based architectures, including Pact-Net[34], USL-Net[35], FAT-Net[36], HTC-Net[37], RM-U-Net[38], ERDUNet[39], MAF-Net[40], DECA-Net[41], CLAS[42], TAM-FCN8s[5], and CoST-UNet[44]. All methods are evaluated on the same datasets, as summarized in Table 4.

Across all datasets, PROMISE-C$_{\text{NN}}$ achieves the highest or near-highest segmentation accuracy while simultaneously attaining the lowest boundary deviation (HD95) and false-negative rate (FNR). On ISIC-2017, it reaches a DSC of 90.7% and an FNR of 7.3%, outperforming HTC-Net by 0.6% in DSC and reducing FNR by approximately 38% (from 11.8% to 7.3%). On CAMUS, PROMISE-C$_{\text{NN}}$ attains a DSC of 91.3% and an IoU of 84.4%, outperforming CC-SAM and TAM-FCN8s, while reducing HD95 to 5.9 pixels, indicating improved boundary precision in low-contrast echocardiographic images. Similar gains are observed on Kvasir-Polyp and Kvasir-Instrument, where PROMISE-C$_{\text{NN}}$ consistently outperforms alternative methods (Table 4). Overall, it yields an average DSC improvement of approximately 2–3% and reduces FNR by up to 35% relative to the strongest competitors, demonstrating that prompt-conditioned channel modulation reliably enhances segmentation accuracy and boundary fidelity across diverse anatomies and imaging modalities.

2.7 Failure Analysis and Limitations

We further analyse failure cases on ISIC-2017 and Kvasir-Polyp (Fig. 7). While PROMISE-Net variants consistently improve contour smoothness and reduce false negatives, residual limitations remain under visually ambiguous or structurally irregular conditions.

In ISIC-2017 (Fig. 7(a)), lesions with internal hollows or heterogeneous pigmentation produce ambiguous intensity gradients that hinder precise boundary localization. Even with prompt-conditioned modulation, internal textures may be conflated with lesion boundaries, leading to mild over-segmentation. This suggests that future designs could benefit from spatial-frequency–aware attention or texture-decoupled feature modulation to better preserve fine internal contrast. Figure 7(b) shows a Kvasir-Polyp failure case in which a single image contains multiple spatially disconnected polyps. Here, a single global prompt or bounding box is insufficient to attend to multiple targets simultaneously, resulting in missed or partially

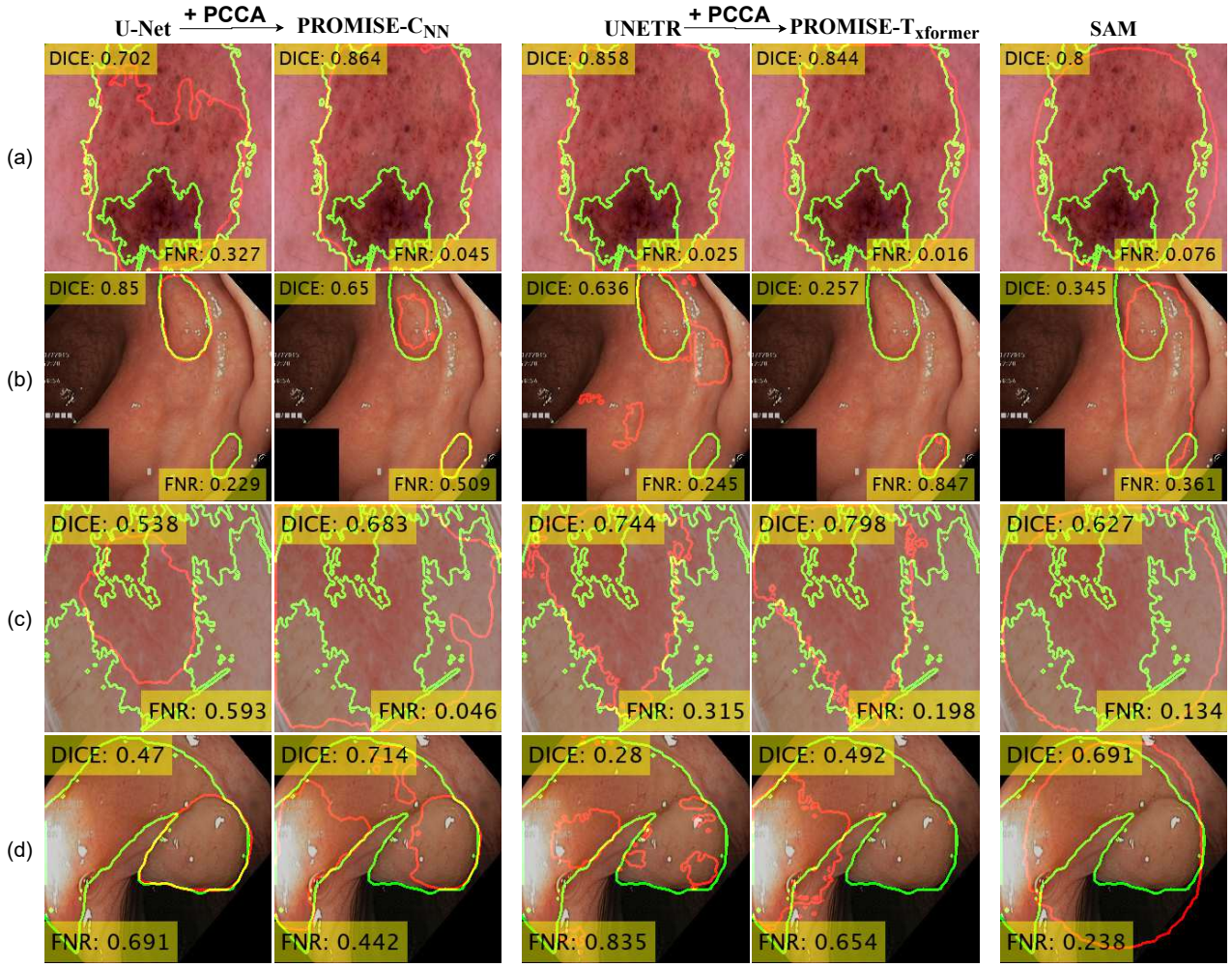


Figure 7. Qualitative analysis of challenging and failure cases on ISIC-2017 and Kvasir-Polyp. Each column compares U-Net, PROMISE-C$_{NN}$, UNETR, PROMISE-T$_{xformer}$, and SAM. Low-contrast textures, disconnected regions, and complex boundary structures lead to degraded performance across all methods; however, PROMISENet variants consistently preserve better boundary conformity than baseline U-Net and UNETR under extreme cases.

segmented regions. More challenging cases in Fig. 7(c)–(d) exhibit highly irregular boundaries, fragmented structures, and specular highlights that introduce strong local ambiguities. Although PROMISE-Net improves boundary adherence relative to baseline and foundation models in these cases, it still struggles to fully capture such complex contours. This motivates future extensions of PROMISE-Net with multi-scale prompt reasoning, adaptive multi-prompting, uncertainty modeling, and richer spatial context aggregation.

3 PROMISE-Net Framework

Building upon the motivation outlined in Section 1, we propose PROMISE-Net, a prompt-aware segmentation framework composed of an image encoder, a prompt encoder, and hierarchical modulation units based on the PCCA module, as illustrated in Fig. 8. At its core, PCCA enables deep, hierarchical integration of semantic prompts within encoder–decoder networks. Specifically, it operates on intermediate feature maps from the image encoder and semantic embeddings from a lightweight prompt encoder to generate spatially modulated representations, wherein PCCA acts as a gating mechanism that regulates the flow of mainstream visual features according to prompt relevance.

The gating behavior of PCCA is analogous to that of an electronic switching transistor: the encoder feature map acts as the primary input (collector current), while the prompt embedding provides the control signal (base drive). This relationship is

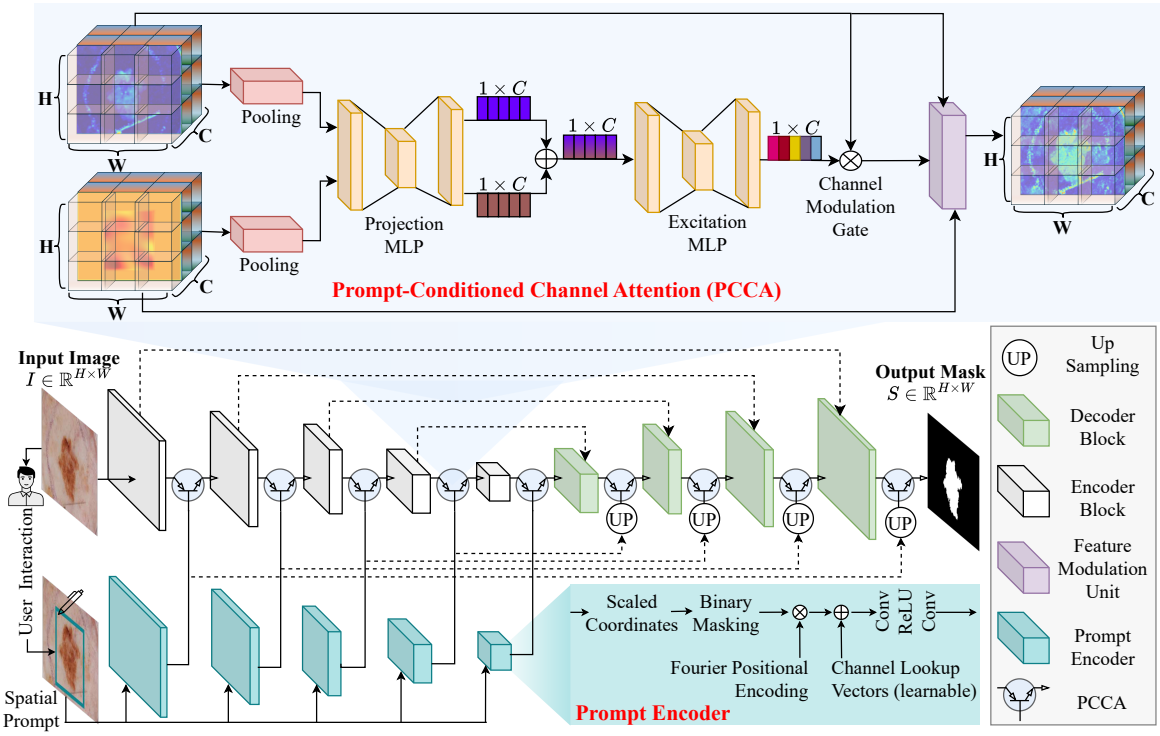


Figure 8. Overview of PROMISE-Net architecture. The top panel presents the proposed PCCA module, while the bottom panel illustrates the segmentation backbone with hierarchical PCCA integration. Embedding PCCA within convolutional blocks yields PROMISE-C_{NN}, while analogous integration in transformer blocks results in PROMISE-$T_{xformer}$.

formalized as:

$$\mathbf{Y} = \sigma(\mathbf{P}) \odot \mathbf{F}, \tag{1}$$

where $\mathbf{F}$ denotes the encoder feature map, $\sigma(\mathbf{P})$ is the prompt-conditioned gate broadcast over $H \times W$, and $\mathbf{Y}$ is the modulated output. When prompt activation is strong, $\sigma(\mathbf{P}) \approx 1$ and the gate conducts; when prompt activation is weak or absent, $\sigma(\mathbf{P}) \approx 0$, suppressing non-salient responses. This transistor-like modulation dynamically amplifies or attenuates spatial activations, thereby focusing the network on anatomically relevant regions.

Compact channel descriptors are extracted from both visual and prompt streams via global average pooling, projected onto a shared latent space, and fused through a gated excitation unit to produce channel-wise modulation weights that adaptively reweight image features at every hierarchical level. Through this design, PCCA preserves spatial semantics while adapting to hierarchical context. PROMISE-Net is designed in two architectural variants: a convolutional model (**PROMISE-C_{NN}**) and a transformer-based model (**PROMISE-$T_{xformer}$**), demonstrating cross-architectural generality. The following subsections describe each component in detail.

3.1 Encoder Representation Learning

Segmentation requires hierarchical feature representations that capture both low-level texture details and high-level semantic cues. To this end, the encoder $\mathscr{E}$ forms the backbone of PROMISE-Net, progressively abstracting visual information into multi-scale latent features that serve as the basis for prompt-conditioned modulation. Given an input image $I^n \in \mathbb{R}^{H \times W}$, the network predicts a segmentation mask $S^n \in \mathbb{R}^{H \times W}$ through the learnable mapping $\mathscr{F}$:

$$S^n = \mathscr{F}(I^n; \theta), \quad n = 1, \ldots, N, \tag{2}$$

where θ denotes the learnable parameters of the segmentation network ($\mathscr{F}$). The encoder extracts progressively higher-level features through convolutional or transformer operations, as shown in (3):

$$\begin{aligned} F^0 &= I^n, \\ F^l &= \sigma\big(\mathrm{BN}(\mathrm{Conv}_{3\times3}(F^{l-1}))\big), \quad l = 1,\ldots,L, \\ F^l &= \mathrm{Down}(F^l), \end{aligned} \tag{3}$$

where $\mathrm{Conv}_{3\times3}$, BN, $\sigma(\cdot)$ and $\mathrm{Down}(\cdot)$ stand for 3×3 convolution, batch normalization, nonlinear activation function (ReLU), and spatial downsampling, respectively. As the network depth increases ($l \to L$), the feature dimensionality expands (64, 128, 256, 512, 1024), enabling the encoder to capture appearance, shape, and boundary cues across multiple spatial scales.

In **PROMISE-C$_{\mathbf{NN}}$**, the hierarchy in (3) is purely convolutional, excelling at capturing local, fine-grained details such as edges and textures. Conversely, **PROMISE-T$_{\mathbf{xformer}}$** replaces convolutional blocks with transformer encoders to model long-range spatial dependencies and enable global contextual reasoning[9], as formalized in (4):

$$F^l = \mathrm{MSA}\big(\mathrm{LN}(F^{l-1})\big) + \mathrm{MLP}\big(\mathrm{LN}(F^{l-1})\big), \tag{4}$$

where MSA, MLP, and LN denote the multi-head self-attention, feed-forward projection, and layer normalization, respectively. Together, these two architectural variants enable the PCCA-coupled PROMISE-Net to evaluate the cross-architectural generality of the proposed modulation mechanism, providing a unified foundation for prompt-aware feature integration.

3.2 Prompt Encoder

The Prompt Encoder transforms a coarse bounding box, provided by the user, into a dense spatial prompt that aligns with the resolution and semantics of the encoder features. Unlike binary masks that merely indicate the region of interest, this design incorporates spatial location, positional context, and semantic embedding, enabling the model to learn both *where* the target region lies and *how* it interacts with the hierarchical feature representations of the backbone.

Given an input bounding box (x_1, y_1, x_2, y_2) defined in image coordinates with spatial dimensions (H_i, W_i), the coordinates are normalized to match the spatial resolution of the corresponding feature map (H_f, W_f) as follows:

$$(x_1', y_1', x_2', y_2') = \left(\tfrac{W_f}{W_i}x_1,\ \tfrac{H_f}{H_i}y_1,\ \tfrac{W_f}{W_i}x_2,\ \tfrac{H_f}{H_i}y_2\right), \tag{5}$$

ensuring that the prompt aligns precisely with the spatial scale of the intermediate feature representation. A binary mask $\mathbf{M}_b \in \{0,1\}^{H_f \times W_f}$ is then generated to localize the region of interest:

$$\mathbf{M}_b(i,j) = \begin{cases} 1, & x_1' \le j < x_2',\ y_1' \le i < y_2', \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

To enrich spatial priors beyond binary localization, we apply a random Fourier positional encoding that maps normalized spatial coordinates $\mathbf{U}(i,j) = [\,j/W_f,\ i/H_f\,]$ into a high-dimensional sinusoidal space, as defined in (7) and inspired by prior work[29,45]:

$$\mathbf{Z}(i,j) = 2\pi\,\mathbf{U}(i,j)\mathbf{G}, \qquad \mathbf{E}(i,j) = [\,\sin\mathbf{Z}(i,j),\ \cos\mathbf{Z}(i,j)\,], \tag{7}$$

where $\mathbf{G} \in \mathbb{R}^{2\times D}$ is a fixed Gaussian projection matrix. The resulting encoding $\mathbf{E} \in \mathbb{R}^{H_f \times W_f \times 2D}$ captures smooth positional variations and introduces global spatial awareness, analogous to positional encodings used in vision transformers[46]. Each bounding box is further assigned a learnable semantic identity through an embedding vector $\mathbf{e}_{\text{box}} \in \mathbb{R}^C$. This embedding is broadcast spatially and fused with the masked positional encoding to produce the preliminary dense prompt representation:

$$\mathbf{F}_b = \big(\mathbf{E} \odot \mathbf{M}_b\big) + \mathbf{e}_{\text{box}} \otimes \mathbf{1}_{H_f \times W_f}, \tag{8}$$

where $\odot$ denotes elementwise multiplication and $\otimes$ represents outer broadcasting. Equation (8) ensures that the encoded prompt emphasizes the target region while introducing a learnable semantic bias that distinguishes bounding-box prompts from other spatial tokens. To enforce local smoothness and reduce box-edge discontinuities, the preliminary prompt representation $\mathbf{F}_b$ is refined through a two-layer convolutional block with shared weights across channels as follows:

$$\hat{\mathbf{F}}_b = \mathrm{Conv}_{3\times3}\Big(\mathrm{ReLU}\big(\mathrm{Conv}_{3\times3}(\mathbf{F}_b)\big)\Big), \tag{9}$$

producing a dense and continuous feature map $\hat{\mathbf{F}}_b \in \mathbb{R}^{C\times H_f\times W_f}$. This refinement acts as a spatial low-pass filter, promoting continuity across bounding-box boundaries and suppressing high-frequency artifacts. For each level l of the encoder-decoder hierarchy with spatial resolution (H_l, W_l) and channel dimension C_l, the refined prompt is projected to a matching scale:

$$\mathbf{P}^l = \mathrm{Resize}\big(\hat{\mathbf{F}}_b\big) \in \mathbb{R}^{C_l\times H_l\times W_l}, \tag{10}$$

where $\mathrm{Resize}(\cdot)$ denotes bilinear interpolation followed by a 1×1 projection ("**UP**" block in Fig. 8). These multi-scale prompt features $\{\mathbf{P}^l\}_{l=1}^{L}$ are subsequently injected into both the encoder and decoder through prompt-conditioned channel attention (Section 3.3).

3.3 Prompt-Conditioned Channel Attention (PCCA)

To integrate spatially embedded priors derived from bounding-box prompts, as described in the previous section, with image features in a semantically consistent and scale-adaptive manner, we introduce the *PCCA* module. Unlike naïve concatenation or additive fusion, PCCA modulates channel activations through prompt-aware excitation, enabling the network to dynamically emphasize channels that are most informative for the user-specified region of interest.

Let the image feature map at stage l be denoted as $\mathbf{F}_b \in \mathbb{R}^{B\times C\times H\times W}$, and let the corresponding prompt embedding be $\mathbf{F}_p \in \mathbb{R}^{B\times D\times H\times W}$. To summarize global contextual information from both streams, pooled descriptors are extracted via global average pooling:

$$\mathbf{v}_b = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{F}_b[:,:,i,j], \qquad \mathbf{v}_p = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{F}_p[:,:,i,j]. \tag{11}$$

Both pooled descriptors are projected into a shared latent space of dimension C using a projection MLP:

$$\tilde{\mathbf{v}}_b = W_b\,\mathbf{v}_b, \qquad \tilde{\mathbf{v}}_p = W_p\,\mathbf{v}_p, \tag{12}$$

where $W_b \in \mathbb{R}^{C\times C}$ and $W_p \in \mathbb{R}^{D\times C}$ are learnable linear transformations that align visual and prompt semantics within a common latent representation. A joint modulation descriptor is then obtained via additive fusion:

$$\mathbf{z} = \tilde{\mathbf{v}}_b + \tilde{\mathbf{v}}_p, \tag{13}$$

which aggregates complementary information from image and prompt cues. The combined descriptor $\mathbf{z}$ is subsequently passed through a lightweight bottleneck gating function implemented as an excitation MLP (inspired by[47]):

$$\mathbf{s} = \sigma\big(W_2\,\delta(W_1\,\mathbf{z})\big). \tag{14}$$

where $W_1 \in \mathbb{R}^{C\times C/r}$ and $W_2 \in \mathbb{R}^{C/r\times C}$ form a two-layer MLP with reduction ratio r, $\delta(\cdot)$ denotes the ReLU activation, and $\sigma(\cdot)$ is the sigmoid gating function.

Equation (14) yields a channel-wise excitation vector $\mathbf{s} \in [0,1]^C$ that adaptively reweights image feature channels according to prompt-conditioned relevance. The modulated feature map is reconstructed as:

$$\mathbf{F}_{\text{out}} = \mathbf{F}_b \odot \mathbf{s} + \mathbf{F}_b \odot \mathbf{F}_p, \tag{15}$$

where $\odot$ denotes elementwise multiplication with broadcasting. Equation (15) encapsulates the "**Feature Modulation Unit**" in Fig. 8, while $\mathbf{F}_b \odot \mathbf{s}$ defines the "**Channel Modulation Gate**". The channel-wise excitation term provides global prompt-conditioned feature selection, while the pixel-wise interaction term complements it by enforcing spatially localized modulation, ensuring that prompt guidance remains both semantically selective and spatially precise.

At inference, the excitation vector $\mathbf{s}$ (in (14)) behaves analogously to the control current of a transistor: when a strong prompt activation $\mathbf{v}_p$ (in (11)) is present, $\mathbf{s}$ opens the gate to amplify feature flow; when prompt activation is weak or absent, $\mathbf{s}$ suppresses non-relevant channels, preventing unwanted activations in unrelated regions. Accordingly, the overall gating behavior of PCCA is compactly expressed as:

$$\mathbf{F}_{\text{out}} = \mathbf{F}_b \odot g(\mathbf{v}_p), \qquad g(\mathbf{v}_p) = \sigma\big(W_2\,\delta\big(W_1\big(W_b\mathbf{v}_b + W_p\mathbf{v}_p\big)\big)\big) + \mathbf{F}_p, \tag{16}$$

where $g(\mathbf{v}_p)$ denotes a prompt-conditioned gating function that combines channel-wise excitation with spatially localized modulation. This generalizes classical squeeze-and-excitation[47] by conditioning the excitation function on both visual and prompt embeddings rather than visual statistics alone.

PCCA modules are embedded at all encoder and decoder levels, ensuring that prompt guidance percolates throughout the network hierarchy. This hierarchical conditioning enables spatially adaptive, semantically consistent, and anatomically plausible feature modulation, allowing the model to maintain focus on clinically relevant structures across scales.

3.4 Prompt-Conditioned Fusion and Decoder Reconstruction

The decoder $\mathscr{D}$ complements the encoder by hierarchically reconstructing spatial details while preserving prompt-guided semantic focus. At each decoder stage, upsampled feature maps are fused with their corresponding encoder features, both already modulated by PCCA, thereby ensuring bidirectional propagation of prompt information throughout the hierarchy.

At each encoder stage l, prompt integration is performed via prompt-conditioned channel attention:

$$\hat{\mathbf{F}}^l = \mathrm{PCCA}\big(\mathbf{F}^l, \mathbf{P}^l\big), \tag{17}$$

where $\hat{\mathbf{F}}^l$ denotes the prompt-modulated encoder feature at level l. During decoding, we denote by $\mathbf{G}^l$ the feature map reconstructed at the l^{th} decoder stage. Prompt information continues to guide feature refinement through successive fusion and modulation operations:

$$\mathbf{G}^L = \hat{\mathbf{F}}^L, \qquad \tilde{\mathbf{G}}^{l-1} = \psi\Big(\mathrm{Up}(\mathbf{G}^l) \| \hat{\mathbf{F}}^{l-1}\Big), \qquad \hat{\mathbf{G}}^{l-1} = \mathrm{PCCA}\big(\tilde{\mathbf{G}}^{l-1}, \mathbf{P}^{l-1}\big), \quad l = L, \ldots, 1. \tag{18}$$

where $\mathrm{Up}(\cdot)$ denotes bilinear upsampling or transposed convolution, and $\psi(\cdot)$ represents a 3×3 convolution followed by normalization and nonlinear activation.

This dual-stage prompt integration enables the encoder and decoder to co-adapt visual evidence and spatial intent, leading to region-specific reconstruction and reduced leakage into irrelevant anatomical areas. The final segmentation mask is obtained:

$$S = \sigma\big(\mathrm{Conv}_{1\times1}(\hat{\mathbf{G}}^0)\big), \tag{19}$$

yielding a dense probability map whose boundaries closely align with the user-indicated regions.

3.5 Unified Forward Formulation and Implementation

Combining prompt-conditioned encoding and reconstruction, the complete PROMISE-Net forward formulation is expressed as:

$$S = \sigma\Big(\mathscr{D}_{\mathrm{PCCA}}\big(\{\hat{\mathbf{F}}^l\}, \{\mathbf{P}^l\}; \theta_{\mathscr{D}}\big)\Big), \qquad \hat{\mathbf{F}}^l = \mathrm{PCCA}\big(\mathscr{E}^l(I; \theta_{\mathscr{E}}), \mathbf{P}^l; \theta_{\mathrm{PCCA}}\big). \tag{20}$$

where prompt-conditioned feature modulation is applied at every encoder stage and propagated through the decoder hierarchy. This formulation captures the bidirectional propagation of spatial priors throughout the encoder-decoder architecture, ensuring that anatomical cues influence both feature abstraction and pixel-level reconstruction. Such dual-stage conditioning enables PROMISE-Net to remain interpretable, responsive to user guidance, and robust across anatomical structures, backbone architectures, and imaging modalities.

From an interpretive standpoint, each PCCA unit can be viewed as a soft gating mechanism that dynamically regulates feature conduction according to prompt activation:

$$\hat{\mathbf{F}}^l = g(\mathbf{P}^l) \odot \mathbf{F}^l, \qquad g(\mathbf{P}^l) = \sigma\big(\mathbf{W}_2^l\, \delta(\mathbf{W}_1^l\, \mathrm{GAP}(\mathbf{P}^l))\big), \tag{21}$$

where $g(\mathbf{P}^l)$ denotes a prompt-dependent conduction gate. When the prompt strongly activates relevant spatial regions, $g(\mathbf{P}^l) \approx 1$ allows unhindered information flow; conversely, when the prompt is weak or absent, $g(\mathbf{P}^l) \approx 0$ suppresses non-salient activations. This transistor-like gating analogy provides an intuitive physical interpretation of PCCA, illustrating how prompt-conditioned modulation adaptively amplifies or attenuates neural feature dynamics across the network hierarchy. The implementation of PROMISE-Net is summarized in Algorithm 1 using PyTorch-style pseudocode.

4 Experimental Datasets and Settings

4.1 Datasets

ISIC-2017 Skin Lesion This dataset[30] comprises 8-bit RGB dermoscopic images with spatial resolutions ranging from 540×722 to 4499×6748 pixels. It includes 2000 training, 150 validation, and 600 test images.

Kvasir-Polyp This dataset[31] consists of 1000 high-quality colonoscopic frames with spatial resolutions ranging from 332×352 to 1920×1072 pixels. It is partitioned into 800 training, 100 validation, and 100 test images.

Kvasir-Instrument This dataset[32] serves as a benchmark for the segmentation of diagnostic and therapeutic instruments in gastrointestinal endoscopy, with spatial resolutions ranging from 571×523 to 1920×1080 pixels. It is partitioned into 472 training, 59 validation, and 59 test images.

Algorithm 1: PROMISE-Net's PyTorch-style pseudocode.

```
# L: number of stages (network's depth) || N: batch size
# (H, W): input size || (H_l, W_l): feature size at stage l
# 𝒟^l: encoder block (3–4) || 𝒟^l: l-th decoder block (17–19)
# Up: upsampling op (bilinear or transposed conv)
# Head: 1×1 conv + sigmoid (19)
# PromptEncoder: (5–10) || PCCA: (11–15)
# Mini-batch: images I, boxes B, and labels Y
for I, B, Y in loader do
    # Prompt encoding (multi-scale)
    F̂_b ← PromptEncoder(B) # (5–9)
    {P^l}_{l=1}^L ← ResizeToPyramid(F̂_b) # (10)
    # Encoder with PCCA and skip collection
    x^0 ← Stem(I) # optional pre-encoding conv
    skips ← [ ]
    for l = 1 to L do
        F^l ← 𝒟^l(x^{l−1}) # (3 or 4)
        F̂^l ← PCCA(F^l, P^l) # (17)
        append(skips, F̂^l)
        x^l ← Down(F̂^l) # stride-2 pooling/conv at l=L
    # Decoder with prompt re-injection via PCCA
    G^L ← F̂^L (bottleneck features)
    for l = L down to 1 do
        U^{l−1} ← Up(G^l)
        S^{l−1} ← skips[l − 1] #prompt-modulated skip
        G̃^{l−1} ← ψ([U^{l−1} || S^{l−1}]) # Conv + BN + ReLU
        Ĝ^{l−1} ← PCCA(G̃^{l−1}, P^{l−1}) # decoder prompting
        G^{l−1} ← Ĝ^{l−1}
    # Prediction head and loss
    Ŝ ← Head(G^0) # 1×1 conv + sigmoid, (19)
    𝒟 ← Loss(Ŝ, Y) # e.g., Dice + BCE, (22)
    # Backward/Update
    optimizer.zero_grad()
    𝒟.backward()
    optimizer.step()
```

CAMUS-Cardiac This dataset[33] comprises echocardiographic scans with heterogeneous image quality, with 35% labeled as good, 46% as medium, and 19% as poor. It includes data from 400 patients for training, 50 for validation, and 50 for testing.

4.2 Training Protocol and Evaluation

Experimental Settings: All experiments were conducted using the PyTorch framework on an NVIDIA Tesla P100 GPU. Images were resized to 256×256 pixels. The models were optimized using the Adam optimizer with a learning rate of 1×10^{-4} and trained for 150 epochs. Extensive offline data augmentation was applied to the CAMUS, Kvasir-Polyp, and Kvasir-Instrument datasets. Specifically, geometric transformations, including horizontal and vertical flips and random rotations of up to 360°, were employed. The augmented data were used during training and validation, while the test sets were kept untouched to ensure unbiased evaluation.

Loss Function: The overall training objective combines multi-class Dice loss with cross-entropy loss to jointly balance structural overlap and pixel-level classification accuracy[48]:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{Dice}} + \lambda\, \mathscr{L}_{\text{CE}}, \tag{22}$$

where $\lambda = 0.5$ in all experiments. The Dice loss is averaged across all foreground classes to emphasize anatomical consistency, while the cross-entropy term stabilizes optimization in the multi-class segmentation setting.

Evaluation: Segmentation accuracy is evaluated using the Dice similarity coefficient (DSC) and Intersection-over-Union

(IoU) for each class. Boundary agreement is quantified using the 95th percentile of the Hausdorff Distance (HD95), which measures the maximum symmetric surface deviation. To further measure pixel-level errors, we report the False Positive Rate (FPR) and False Negative Rate (FNR), capturing over- and under-segmentation, respectively. In addition to these spatial metrics, for cardiac datasets such as CAMUS, we assess the ejection fraction (EF)[49], a clinically relevant functional biomarker defined as the ratio of stroke volume to end-diastolic volume (EDV). Agreement between estimated and reference EF values is analysed using Bland–Altman plots, which report the mean bias and limits of agreement (LoA) across the test cohort.

Data Availability

This study used four publicly available datasets. The ISIC-2017 Skin Lesion dataset, comprising dermoscopic images for skin lesion segmentation, is available at the ISIC Challenge Archive (https://challenge.isic-archive.com/data/#2017). The Kvasir-SEG dataset, consisting of gastrointestinal polyp images with corresponding segmentation masks, is available at https://datasets.simula.no/kvasir-seg/. The Kvasir-Instrument dataset, containing endoscopic images of gastrointestinal surgical instruments, is available at https://datasets.simula.no/kvasir-instrument/. The CAMUS dataset, comprising 2D echocardiographic images for cardiac structure segmentation, is available at the CREATIS challenge portal (https://www.creatis.insa-lyon.fr/Challenge/camus/).

Author contributions statement

M.H. and **M.K.H.** conceived the study, developed the methodology, implemented the software, and performed the experiments. **M.H.**, **M.R.I.**, and **M.K.H.** conducted the formal analysis, carried out the investigation, and wrote the original draft of the manuscript. **M.H.** and **L.H.** curated the data and performed validation. **E.S.** and **M.K.H.** supervised the project. All authors reviewed and approved the manuscript.

Additional information

Competing interests: The authors declare no competing interests.
Funding: Not applicable

References

1. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Analysis* **42**, 60–88, DOI: https://doi.org/10.1016/j.media.2017.07.005 (2017).
2. Zhang, J., Xie, Y., Xia, Y. & Shen, C. Attention residual learning for skin lesion classification. *IEEE Transactions on Med. Imaging* **38**, 2092–2103, DOI: 10.1109/TMI.2019.2893944 (2019).
3. Gutman, D. *et al.* Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1605.01397* (2016).
4. Hasan, M. K. *et al.* Dermo-doctor: A framework for concurrent skin lesion detection and recognition using a deep convolutional neural network with end-to-end dual encoders. *Biomed. Signal Process. Control.* **68**, 102661 (2021).
5. Hasan, M. K., Yang, G. & Yap, C. H. Motion-enhanced cardiac anatomy segmentation via an insertable temporal attention module. In *International Workshop on Advances in Simplifying Medical Ultrasound*, 143–153 (Springer, 2025).
6. Wang, R., Zhou, Y., Zhao, C. & Wu, H. A hybrid flower pollination algorithm based modified randomized location for multi-threshold medical image segmentation. *Bio-medical Mater. Eng.* **26**, S1345–S1351 (2015).
7. Patil, D. D. & Deore, S. G. Medical image segmentation: a review. *Int. J. Comput. Sci. Mob. Comput.* **2**, 22–27 (2013).
8. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241 (Springer, 2015).
9. Hatamizadeh, A. *et al.* Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584 (2022).
10. Chen, X. *et al.* Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Analysis* **79**, 102444 (2022).
11. Zhang, X. *et al.* An anatomy-and topology-preserving framework for coronary artery segmentation. *IEEE Transactions on Med. Imaging* **43**, 723–733 (2023).

12. Chen, Z., Tian, Z., Zhu, J., Li, C. & Du, S. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11676–11685 (2022).

13. Kuang, Z., Yan, Z., Zhou, H. & Yu, L. Cluster-re-supervision: Bridging the gap between image-level and pixel-wise labels for weakly supervised medical image segmentation. *IEEE J. Biomed. Heal. Informatics* **27**, 4890–4901 (2023).

14. Zhai, S. *et al.* Pa-seg: Learning from point annotations for 3d medical image segmentation using contextual regularization and cross knowledge distillation. *IEEE Transactions on Med. Imaging* **42**, 2235–2246 (2023).

15. Wei, J., Hu, Y., Cui, S., Zhou, S. K. & Li, Z. Weakpolyp: You only look bounding box for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 757–766 (Springer, 2023).

16. Zhang, K. & Zhuang, X. Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11656–11665 (2022).

17. Li, Z. *et al.* Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation. *IEEE Transactions on Med. Imaging* **43**, 2254–2265 (2024).

18. Valvano, G., Leo, A. & Tsaftaris, S. A. Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Med. Imaging* **40**, 1990–2001 (2021).

19. Zhang, K. & Zhuang, X. Shapepu: A new pu learning framework regularized by global consistency for scribble supervised cardiac segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 162–172 (Springer, 2022).

20. Tajbakhsh, N. *et al.* Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med. Image Analysis* **63**, 101693 (2020).

21. Luo, X. *et al.* Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 528–538 (Springer, 2022).

22. Wang, A., Xu, M., Zhang, Y., Islam, M. & Ren, H. S 2 me: Spatial-spectral mutual teaching and ensemble learning for scribble-supervised polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 35–45 (Springer, 2023).

23. Lei, Y., Luo, H., Wang, L., Zhang, Z. & Zhang, L. Pclmix: Weakly supervised medical image segmentation via pixel-level contrastive learning and dynamic mix augmentation. In *International Conference on Intelligent Computing*, 62–73 (Springer, 2024).

24. Wu, H., Li, X., Lin, Y. & Cheng, K.-T. Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation. *IEEE Transactions on Med. Imaging* **42**, 3244–3255 (2023).

25. Han, M. *et al.* Scribble-based 3d multiple abdominal organ segmentation via triple-branch multi-dilated network with pixel-and class-wise consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 33–42 (Springer, 2023).

26. Yang, Z., Lin, D., Ni, D. & Wang, Y. Non-iterative scribble-supervised learning with pacing pseudo-masks for medical image segmentation. *Expert. Syst. with Appl.* **238**, 122024 (2024).

27. Zhang, X., Zhu, L., He, H., Jin, L. & Lu, Y. Scribble hides class: Promoting scribble-based weakly-supervised semantic segmentation with its class label. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 7332–7340 (2024).

28. Wang, Y., Peng, J. & Zhang, Z. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9092–9101 (2021).

29. Kirillov, A. *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026 (2023).

30. Codella, N. C. F. *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). *CoRR* **abs/1710.05006** (2017). 1710.05006.

31. Jha, D. *et al.* Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 451–462 (Springer, 2020).

32. Jha, D. *et al.* Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling*, 218–229 (Springer International Publishing, Cham, 2021).

33. Leclerc, S. *et al.* Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Med. Imaging* **38**, 2198–2210, DOI: 10.1109/TMI.2019.2900516 (2019).

34. Chen, W. *et al.* Pact-Net: Parallel cnns and transformers for medical image segmentation. *Comput. Methods Programs Biomed.* **242**, 107782 (2023).

35. Li, X. *et al.* USL-Net: Uncertainty self-learning network for unsupervised skin lesion segmentation. *Biomed. Signal Process. Control.* **89**, 105769 (2024).

36. Wu, H. *et al.* FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Med. Image Analysis* **76**, 102327 (2022).

37. Tang, H. *et al.* HTC-Net: A hybrid cnn-transformer framework for medical image segmentation. *Biomed. Signal Process. Control.* **88**, 105605 (2024).

38. Tang, H. *et al.* RM-UNet: UNet-like Mamba with rotational SSM module for medical image segmentation. *Signal, Image Video Process.* **18**, 8427–8443 (2024).

39. Li, H., Zhai, D.-H. & Xia, Y. ERDUnet: An efficient residual double-coding unet for medical image segmentation. *IEEE Transactions on Circuits Syst. for Video Technol.* **34**, 2083–2096 (2023).

40. Yang, L., Gu, Y., Bian, G. & Liu, Y. MAF-Net: A multi-scale attention fusion network for automatic surgical instrument segmentation. *Biomed. Signal Process. Control.* **85**, 104912 (2023).

41. Liang, S., Zhang, J., Bian, A. & You, J. DECA-Net: Dual encoder and cross-attention fusion network for surgical instrument segmentation. *Pattern Recognit. Lett.* **185**, 130–136 (2024).

42. Wei, H. *et al.* Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In *MICCAI 2020*, 623–632 (Springer International Publishing, Cham, 2020).

43. Li, X. *et al.* EchoSAM: SAM adaption for unified 2D echocardiography segmentation and ejection fraction calculation. *Biomed. Signal Process. Control.* **109**, 108000, DOI: https://doi.org/10.1016/j.bspc.2025.108000 (2025).

44. Islam, M. R., Qaraqe, M. & Serpedin, E. CoST-UNet: Convolution and swin transformer based deep learning architecture for cardiac segmentation. *Biomed. Signal Process. Control.* **96**, 106633, DOI: https://doi.org/10.1016/j.bspc.2024.106633 (2024).

45. Tancik, M. *et al.* Fourier features let networks learn high frequency functions in low dimensional domains. *Adv. Neural Inf. Process. Syst.* **33**, 7537–7547 (2020).

46. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

47. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141 (2018).

48. Hasan, M. K., Dahal, L., Samarakoon, P. N., Tushar, F. I. & Marti, R. DSNet: Automatic dermoscopic skin lesion segmentation. *Comput. Biol. Medicine* **120**, 103738 (2020).

49. Hasan, M. K., Zhu, H., Yang, G. & Yap, C. H. Deep learning image registration for cardiac motion estimation in adult and fetal echocardiography via a focus on anatomic plausibility and texture quality of warped image. *Comput. Biol. Medicine* **187**, 109719 (2025).