Towards Noise-Robust Medical Segmentation via Chebyshev-Attention-Based Asymmetric UNet

Anonymous WACV Algorithms Track submission

Paper ID 2274

Abstract

*Existing medical image segmentation methods based on UNet architectures exhibit significant noise sensitivity due to cascaded feature propagation in symmetric encoder-decoder paths and linear feature fusion mechanisms. To address this, we propose **CASUNet** (**C**hebyshev-**A**ttention-Based A**s**ymmetric UNet), a noise-resilient framework integrating an asymmetric **UNet** backbone with a novel **CPA** (**C**hebyshev **P**olynomial **A**ggregation) module. Specifically, CASUNet aggregates hierarchical representations by grouping multi-scale features into distinct low- and high-resolution branches rather than sequentially upsampling, effectively mitigating noise sensitivity through isolated feature processing. Furthermore, CASUNet introduces the CPA mechanism where hierarchical features are expanded into orthogonal polynomial terms, enhancing feature fusion capacity beyond linear aggregation while ensuring noise robustness through a carefully designed polynomial-normalization. Theoretical analysis establishes desirable properties of the proposed model, while extensive experiments verify its superior noise immunity and competitive performance compared to state-of-the-art approaches with significantly enhanced parameter efficiency.*

1. Introduction

Medical image segmentation plays a pivotal role in clinical workflows, enabling precise localization of anatomical structures and pathological regions for disease diagnosis, treatment planning, and surgical guidance. Among various applications, polyp segmentation in colonoscopy images is critical for early detection of colorectal cancer, where a 1% improvement in polyp detection reduces cancer risk by 3%. While U-Net [27] and its CNN variants [12, 17, 18, 20, 25, 33, 36] excel at local feature extraction, their limited receptive fields hinder long-range dependency modeling. Vision Transformers (ViTs) [10] address this via global self-attention in models like Polyp-PVT [9], DuAT [31], and SSFormer [28], yet often sacrifice fine-grained details, causing blurred boundaries. Hybrid approaches synergize CNNs and ViTs through multi-scale fusion, such as Meta-Polyp [32], FCB-SwinV2 [13], and TransUNet [5].

Medical image segmentation remains critically vulnerable to noise artifacts, including Gaussian [14, 30], Poisson [30], Rician [7], and physical artifact noise [3], directly compromising diagnostic accuracy in low-dose CT, MRI, and ultrasound. Recent approaches including anatomical anchor alignment frameworks like A3-DualUD [35] and diffusion-based denoisers such as FDiff-Fusion [8], face persistent challenges. These include over-smoothing of fine anatomical details like WSSS [24], and prohibitive computational latency from iterative denoising processes [4]. Even foundation models such as SAMed-2 [37] further inherit limitations in handling rare pathologies despite noise-suppression mechanisms. The performance of models on noisy images are shown in Figure 1.

Based on this, we further propose that the symmetric encoder-decoder structure of UNet is the root cause of significant noise sensitivity as demonstrated in Theorem 1, based on which, we propose **CASUNet (Chebyshev-Attention-Based Asymmetric UNet)**, a novel segmentation framework built upon a proposed well-designed **UNet** backbone with integrated **Chebyshev Polynomial Aggregation (CPA)** module. Specifically, our design deconstructs the symmetric encoder-decoder by discarding traditional upsampling paths to truncate error-prone upsampling paths, and grouping multi-scale features into isolated low/high-resolution branches to prevent cross-scale contamination, significantly suppressing noise amplification. Then the grouped hierarchical features are aggregated by CPA instead of conventional linear aggregation to enhance cross-scale interaction and further improve noise robustness through Chebyshev polynomial expansion.

Our contributions can be summarized as follows:

1. We rigorously prove that standard UNet architectures exhibit high noise amplification due to cascaded feature

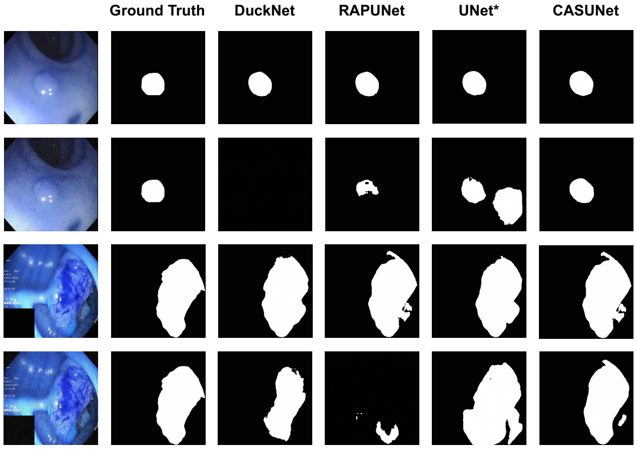


Figure 1. Some segmentation cases: rows 1 and 3 show the noise-free cases, while rows 2 and 4 show predictions after adding noise to the images in the preceding rows. Existing high-performing networks struggle to correctly segment noisy images. By contrast, our proposed CASUNet achieves segmentation accuracy on noisy images that is comparable to their accuracy on noise-free images.

propagation in the decoder.

2. We propose an asymmetric UNet backbone, which preserves UNet's downsampling path but replaces the upsampling path with a hierarchical feature grouping mechanism followed by aggregation. This design significantly reduces noise sensitivity compared to traditional UNet structures.
3. We introduce Chebyshev Polynomial Aggregation (CPA), enhancing cross-level feature interactions while surpassing linear aggregation in noise robustness through a carefully designed polynomial-normalization scheme.

To comprehensively evaluate the effectiveness of CASUNet, we first conduct extensive experiments across multiple polyp segmentation benchmarks, including Kvasir-SEG [19], CVC-ClinicDB [2], CVC-ColonDB [1], and ETIS-LaribPolypDB [29]. Then we add noise to images to evaluate the noise robustness of the proposed model. Last we conduct ablation studies on the structure and the polynomial order. Our experiments demonstrate that CASUNet achieves SOTA-comparable performance and stronger noise robustness while reducing parameters and FLOPs.

2. Related Work

2.1. Methods for Medical Image Segmentation

Recent advancements in medical image segmentation have witnessed a progression from CNN-based architectures to transformer-based models and their hybrids, each addressing distinct challenges in feature extraction and contextual modeling. Since the introduction of U-Net [27], CNNs have been foundational, with variants like PraNet [12] introducing reverse attention for refining uncertain regions, ResUNet++ [20] employing residual blocks for improved local detail extraction, and Double UNet [21] integrating ASPP [6] and Squeeze-and-Excitation [16] modules to capture global dependencies. Despite their success, CNNs remain limited in modeling long-range contextual relationships due to their inherently local receptive fields. Vision Transformers (ViTs) [10] revolutionized global feature integration through self-attention mechanisms, with methods like Polyp-PVT [9] using cascaded fusion modules (CFM/CIM/SAM) to balance local and global features and SSFormer [28] employing a progressive locality decoder for stepwise aggregation. However, transformer-based models often struggle with boundary detail extraction due to their patch-based operations. Hybrid architectures combine the strengths of both paradigms. For instance, Meta-Polyp [32] integrates MetaFormer [34] with multi-scale upsampling, TransUNet [5] fuses ViT's global context with CNN decoders, and FCB-SwinV2 [13] advance parallel CNN-transformer processing. DuckNet [11] and RAPUNet [22] further enhance receptive field capabilities via atrous convolutions, achieving state-of-the-art performance but increasing computational complexity.

2.2. Noise and Denoising Methods

Medical image noise manifests in modality-dependent forms, such as Gaussian in CT/MRI [14, 30], Poisson in low-dose imaging [30], Rician in MRI [7], and physical artifacts like motion/beam-hardening [3], each distorting anatomical boundaries in distinct ways. Critically, this heterogeneity renders universal denoising impossible, and recent methods exhibit fundamental flaws. Anatomical anchor alignment (A3-DualUD [35]) and mutual information maximization (SFDA-MIM [26]) ignore spatially varying noise patterns, and WSSS [24] causes boundary oversmoothing of fine-grained details. Besides, diffusion-based denoisers (FDiff-Fusion [8]) suffer latency from iterative processes [4]. Even foundation models (SAMed-2 [37]) lack robustness to rare pathologies.

3. Notations and Assumptions

Let x_k, y_k denote the feature maps of the k-th layer of the downsampling and upsampling process, respectively. We use $\Delta\cdot$ to indicate the noise of a given tensor. Specifically, $\Delta\boldsymbol{x} \in \mathbb{R}^{d_{in}}$ and $\Delta\boldsymbol{y} \in \mathbb{R}^{d_{out}}$ serve as an additive noise vector corrupting the input and the propagated output noise vector. Generally, we use d_{in} and d_{out} to stand for the input and output dimensions. f_k, g_k indicate the k-th layer of the network function in downsampling and upsampling, although the notation $f : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ stands for a general layer for simplicity. And $\boldsymbol{J} \in \mathbb{R}^{d_{out}\times d_{in}}$ indicates the

Layer	Conv	Trans Conv	Max Pool	ReLU	Linear Attn
EEA	2	$2st^2$	$1/h^2$	0.5	$2d_{out}/d_{in}$

Table 1. The expected energy amplification factor (EEA) of various network layers, where st, h indicates the stride of a convolutional layer and a max-pooling layer.

Jacobian matrix of the layer f, with entries $J_{ij} = \partial f_i/\partial x_j$. Besides, s denotes the interaction of the input features while u serves as the output of the aggregation module.

When it does not cause ambiguity, we do not distinguish between low- and high-level vectors and instead use the symbol with the subscript $\cdot_l$ or $\cdot_h$ omitted.

The basic assumptions and the corresponding justifications are listed as follows.

1. **Noise Model** We model the noise $\Delta\boldsymbol{x}$ as zero-mean white noise, indicating that its components are independent and identically distributed with variance σ^2. Its statistical properties are

$$\mathbb{E}[\Delta\boldsymbol{x}] = \boldsymbol{0}, \quad \text{Cov}(\Delta\boldsymbol{x}) = \mathbb{E}[\Delta\boldsymbol{x}\Delta\boldsymbol{x}^\top] = \sigma^2\boldsymbol{I}_{d_{in}}$$

   Justification: This is a standard and widely-used noise model for thermal and sensor noise.

2. **Local Linearity** We analyze noise propagation using a first-order Taylor approximation of the layer function, formalized as $f(\boldsymbol{x} + \Delta\boldsymbol{x}) = f(\boldsymbol{x}) + \boldsymbol{J}\Delta\boldsymbol{x} + O(\|\Delta\boldsymbol{x}\|^2) \approx f(\boldsymbol{x}) + \boldsymbol{J}\Delta\boldsymbol{x}$. The output noise is thus $\Delta\boldsymbol{y} \approx \boldsymbol{J}\Delta\boldsymbol{x}$.
   Justification: This approximation is accurate for small perturbations, which is the standard assumption in sensitivity and noise propagation analysis in deep networks.

3. **Weight Initialization.** We assume the convolutional weights are initialized using the Kaiming initialization scheme [15]. For a weight tensor with a fan-in of n_{in}, the weights w are drawn from a distribution with $\text{Var}(w) = 2/n_{in}$.
   Justification: This is the standard operation for initializing deep networks with ReLU activations to ensure stable gradient propagation during training.

4. Method

We formally define and rigorously establish the presence of noise amplification in the UNet architecture in Section 4.1, serving as the core motivation of our method. Then we introduce CASUNet, a noise-resilient medical image segmentation framework built upon an asymmetric UNet backbone and Chebyshev Polynomial Aggregation (CPA) module, as illustrated in Figure 2 and Figure 3c. Section 4.2 details the structural design of the asymmetric UNet. We then present the CPA module in Section 4.3. The detailed proof of the mentioned theorems and propositions is illustrated in Appendix A.

4.1. Noise in UNet Architectures

We analyze the noise in UNet architectures from the perspective of energy, for which we define the Expected Energy Amplification Factor (EEA) $\mathcal{A}$ as the ratio of the expected output energy to the expected input energy.

The expected energy of the input noise is

$$\mathbb{E}\left[\|\Delta\boldsymbol{x}\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^{d_{in}}(\Delta x_i)^2\right] = \sum_{i=1}^{d_{in}}\mathbb{E}[(\Delta x_i)^2] = d_{in}\sigma^2$$

The expected energy of the output noise is calculated as shown in Eq (1)

$$\begin{aligned}\mathbb{E}[\|\Delta\boldsymbol{y}\|_2^2] &= \mathbb{E}[\|\boldsymbol{J}\Delta\boldsymbol{x}\|_2^2] = \mathbb{E}[(\boldsymbol{J}\Delta\boldsymbol{x})^\top(\boldsymbol{J}\Delta\boldsymbol{x})] \\ &= \mathbb{E}[\Delta\boldsymbol{x}^\top\boldsymbol{J}^\top\boldsymbol{J}\Delta\boldsymbol{x}] = \text{Tr}(\boldsymbol{J}^\top\boldsymbol{J}\cdot\mathbb{E}[\Delta\boldsymbol{x}\Delta\boldsymbol{x}^\top]) \\ &= \text{Tr}(\boldsymbol{J}^\top\boldsymbol{J}\cdot\sigma^2\boldsymbol{I}_{d_{in}}) = \sigma^2\text{Tr}(\boldsymbol{J}^\top\boldsymbol{J}) \\ &= \sigma^2\|\boldsymbol{J}\|_F^2 \qquad (1)\end{aligned}$$

Thus, the amplification factor $\mathcal{A}$ can be computed according to its definition, as shown in Eq (2).

$$\mathcal{A} = \frac{\mathbb{E}[\|\Delta\boldsymbol{y}\|_2^2]}{\mathbb{E}[\|\Delta\boldsymbol{x}\|_2^2]} = \frac{\sigma^2\|\boldsymbol{J}\|_F^2}{d_{in}\sigma^2} = \frac{\|\boldsymbol{J}\|_F^2}{d_{in}} \qquad (2)$$

A layer is expected to amplify noise energy if $\mathcal{A} > 1$, which is equivalent to the condition $\|\boldsymbol{J}\|_F^2 > d_{in}$. This condition signifies that the average squared singular value of the Jacobian is greater than one.

Building upon this, we propose Theorem 1 as follows.

Theorem 1 (The noise is amplified during the propagation in UNet architectures)**.** *A standard UNet architecture whose layers are initialized using a modern variance-preserving scheme (e.g., Kaiming initialization) is expected to amplify the energy of additive white noise. The amplification is particularly pronounced in the decoder path due to the expansive nature of transposed convolutions.*

The detailed proof is provided in Appendix A.1, during which we analyze the EEA of primary layers in a UNet, including the convolution, transposed convolution, max pooling, ReLU, and linear attention layers, as shown in Table 1.

Theorem 1 demonstrates that it is the symmetric encoder-decoder structure, especially the upsampling of UNet, leading to significant noise sensitivity.

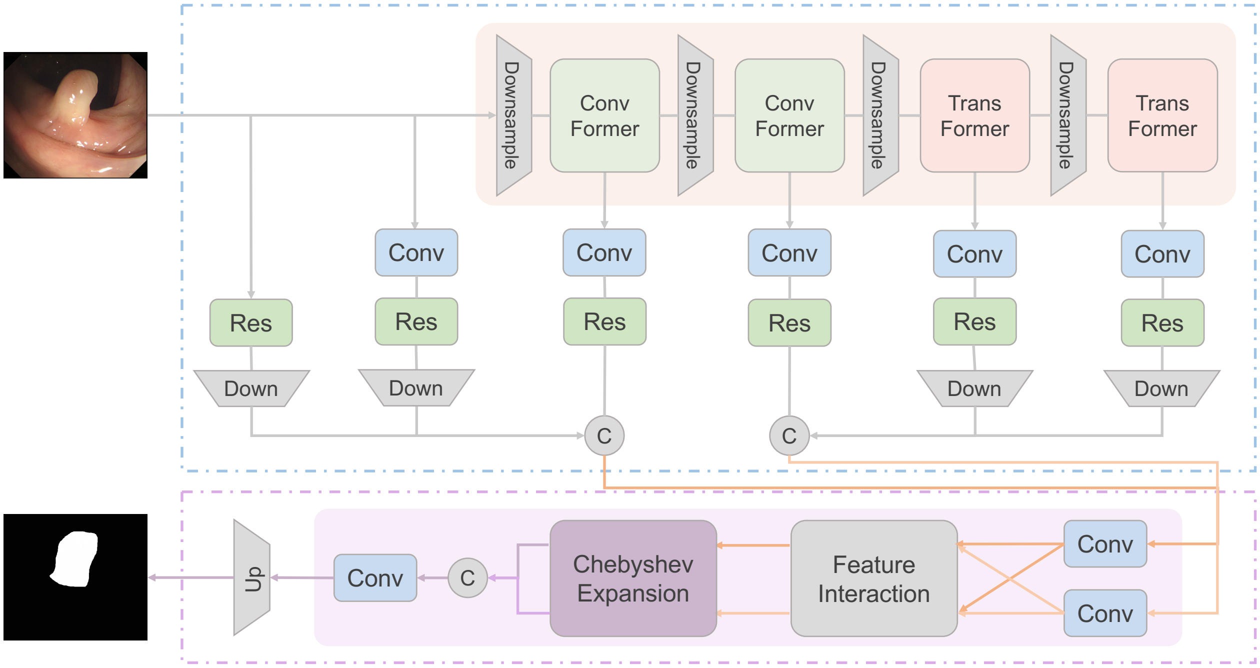


Figure 2. Left: The overall architecture of CASUNet. The top blue dashed box and the bottom pink dashed box respectively denote the encoder and decoder parts. The orange shaded box represents a 4-layer CAFormer, and the purple shaded box denotes the CPA module. And the blue box represents a standard 1×1 convolutional layer, the green box indicates a residual block consisting of two 3×3 convolutional layers.

4.2. The Asymmetric UNet Structure

To mitigate the noise amplification inherent in the standard UNet, a straightforward solution is to modify the conventional decoder to avoid extensive cascaded upsampling.

Based on the rationale, we construct an asymmetric UNet architecture, as illustrated in Figure 2. Our framework consists of two key components: a hybrid CNN-transformer encoder and an asymmetric feature grouping decoder. The encoder retains the downsampling path of traditional UNet, leveraging CAFormer, a variant of MetaFormer [34] for global contextual modeling and two-layer-convolutional ResNet blocks for local texture refinement, enclosed in blue and pink dashed boxes, respectively. The decoder discards symmetric upsampling paths and instead groups multi-scale features into isolated low/high-resolution branches, which are fused through the following Chebyshev Polynomial Aggregation (CPA) module. This design fundamentally reduces noise amplification by truncating error-prone gradient propagation while preserving cross-scale interaction.

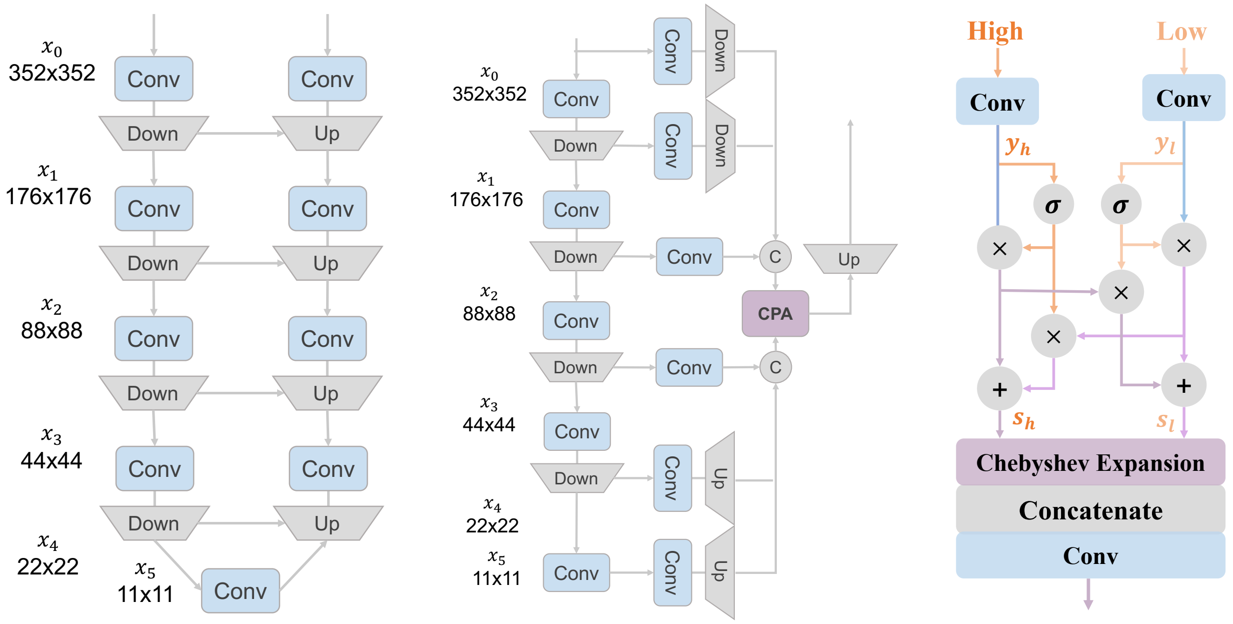
4.2.1. Hybrid CNN-Transformer Encoder

The encoder integrates a pretrained CAFormerS18 [34] backbone with two-layer convolutional ResNet blocks to balance global and local feature extraction, which is a four-stage architecture combining the strengths of CNNs and ViTs by employing depthwise separable convolutions in the first two stages and self-attention modules in the latter two. This hybrid design naturally integrates the local feature extraction capability of CNNs with the global modeling power of transformers, making it an ideal backbone for our task.

As shown in Figure 2 and Figure 3b, the input tensor passing through a shape-preserving convolutional layer, along with the outputs from each of the four stages of CAFormer is downsampled using a stride-2 convolution followed by a two-layer ResNet block. Assuming an initial input resolution of 352×352, the resulting feature maps have spatial dimensions ranging from 352×352 down to 11×11. This hierarchical downsampling enables a robust fusion of features from both the CNN-based and transformer-based components, facilitating rich multi-scale representation learning.

4.2.2. Hierarchical Feature Grouping Decoder

Instead of adopting the conventional UNet upsampling structure that symmetrically mirrors the encoder, we propose a novel feature aggregation strategy by grouping and combining features from different levels of the encoder. Specifically, the three highest-level features are first passed through individual two-layer ResNet blocks and then concatenated to form a high-level feature map that captures fine-grained image details. Similarly, the three lowest-level features are processed in the same manner to produce a low-level feature map representing global structural infor-



(a) UNet (b) CASUNet (c) CPA

Figure 3. The structure of (a) UNet, (b) CASUNet, and (c) the CPA module.

mation. The structure of CASUNet is exhibited in Figure 3b compared to UNet shown in Figure 3a.

The high-level and low-level feature maps are then fed into our Chebyshev Polynomial Aggregation (CPA) module (introduced in Section 4.3) for further fusion. We find that **this Asymmetric UNet architecture exhibits greater robustness to noise compared to the standard UNet structure,** as illustrated in Proposition 1:

Proposition 1. *The proposed asymmetric UNet architecture achieves at least 30% reduction in EEA compared to the standard UNet.*

where the detailed proof can be found in Appendix A.2.

As illustrated in Proposition 1, our design avoids the gradient multiplication caused by multiple cascaded layers, thereby enhancing robustness to noise compared to the typical UNet structure.

4.3. Chebyshev Polynomial Aggregation

As shown in Figure 3c, the Chebyshev Polynomial Aggregation (CPA) module extends linear feature fusion to orthogonal polynomial dynamics. Specifically, the CPA module takes low-level and high-level features as inputs, both of which are first passed through cascaded convolutional layers followed by a sigmoid function.

Taking the low-level feature y_l as an example, we use the sigmoid-activated vector $\sigma(y_l)$ as the weight for the low-level feature itself, applying element-wise multiplication to achieve feature enhancement. Meanwhile, we use $(1-\sigma(y_l))$ as the weight for the high-level feature y_h, after it has been multiplied element-wise with its own sigmoid-activated version $\sigma(y_h)$. The weighted combination of these two enhanced features forms the interaction term s_l. Similarly, we derive the interaction term s_h for the high-level feature. The linear combination process is shown in Eq (3). Next, there are various aggregation processes for this interaction feature s. First, it can be directly added to the input feature y. Second, it can be expanded into vanilla higher-order polynomials and summed up. We set these methods as baselines and introduce our Chebyshev polynomial aggregation, i.e., the interaction term is expanded into higher-order Chebyshev polynomial terms and summed up. According to the processes, we name these methods as Linear Aggregation (LA), Polynomial Aggregation (PA), and Chebyshev Polynomial Aggregation (CPA), as exhibited in Eq (5). Finally, the processed high- and low-level features are concatenated together and passed through a convolutional layer to produce the final output.

$$\begin{cases} s_l = \sigma(y_l)y_l + (1-\sigma(y_l))\sigma(y_h)y_h \\ s_h = \sigma(y_h)y_h + (1-\sigma(y_h))\sigma(y_l)y_l \end{cases} \tag{3}$$

$$T^{(k)}(s) = \begin{cases} 0, & k=0 \\ s, & k=1 \\ 2sT^{(k-1)}(s) - T^{(k-2)}(s), k>1 \end{cases} \tag{4}$$

$$u = \begin{cases} y + \sum_{k=1}^{d} T^{(k)}(s) & \text{(CPA)} \\ y + s & \text{(LA)} \\ y + \sum_{k=1}^{d} s^k & \text{(PA)} \end{cases} \tag{5}$$

Datasets			**Kvasir-SEG**				**CVC-ClinicDB**			
Models	Params(M)	FLOPS(G)	mDice	mIoU	mPrec	mRec	mDice	mIoU	mPrec	mRec
PraNet	32.5	52.0	0.898	0.840	N/A	N/A	0.899	0.849	N/A	N/A
ResUNet++	**14.5**	71.0	0.813	0.793	0.706	0.877	0.796	0.796	0.702	0.879
FCBFormer	53.0	149.5	0.939	**0.890**	0.946	**0.940**	0.947	0.902	0.953	0.944
DUCKNet	592.8	365.3	0.905	0.883	0.910	0.872	0.948	0.901	0.947	0.949
RAPUNet	43.8	42.9	0.924	0.859	0.939	0.910	0.952	0.912	0.949	**0.959**
UNet*	35.0	44.1	0.901	0.819	0.932	0.872	0.950	0.905	0.948	0.942
CASUNet-LA	31.2	**31.1**	0.923	0.857	**0.965**	0.886	0.952	0.910	0.951	0.942
CASUNet-CPA-o5	31.2	**31.1**	**0.939**	0.885	0.942	0.936	**0.954**	**0.912**	**0.955**	0.953

Table 2. The segmentation performance on Kvasir-SEG and CVC-ClinicDB, where CASUNet-LA and CASUNet-CPA-o5 indicate the Asymmetric Unet architecture with linear aggregation module and Chebyshev Polynomial Aggregation module with 5 order, respectively

We find that polynomial aggregation, compared to linear aggregation, can enhance the information interaction from the perspective of the Hilbert-Schmidt Independence Criterion (HSIC) [23], a kernel-based statistic that quantifies dependence between two random variables by computing the Hilbert–Schmidt norm of their cross-covariance operator in Reproducing Kernel Hilbert Spaces (RKHS). To formalize this observation, we present the following proposition:

Proposition 2 (The interaction gain of polynomial aggregation)**.** *Assuming s_l, s_h as the low- and high-level interaction term, the interaction gain of polynomial aggregation compared to linear aggregation is:*

$$\Delta HSIC = HSIC_{PA} - HSIC_{LA} \geq \sum_{k=2}^{n} \|C_{s_l^k s_h^k}\|_{HS}^2 > 0$$

We find that not only can the Chebyshev polynomial aggregation improve interaction gain, but it can also achieve lower noise amplification compared to linear aggregation applied with a simple polynomial normalization. We formalize this conclusion in the following proposition.

Proposition 3 (The noise of Chebyshev polynomial aggregation with polynomial normalization)**.** *Under the Chebyshev polynomial aggregation with polynomial normalization, the upper bound of the noise in the aggregated features is given by $\left[(\frac{\pi}{2}+1)\nabla_y s + \mathbb{I}\right]\varepsilon$, where ε indicates the noise along with y.*

As shown in Proposition 3, the upper bound of the noise introduced by Chebyshev polynomial aggregation can be restricted lower than that of the linear aggregation, denoted as $\nabla_y s \cdot \varepsilon$.

We should claim that it is the unique construction of the Chebyshev polynomial that restricts the noise, because general polynomial aggregation will introduce systematically higher noise than linear aggregation, as shown in Proposition 4:

Proposition 4 (The noise of general polynomial aggregation)**.** *Let ε denote the noise of the input feature y, then the noise is amplified to $\left[\sum_{i=1}^{d} i s^{i-1}\nabla_y s + \mathbb{I}\right]\cdot\varepsilon$.*

According to Proposition 4, the noise of general polynomial aggregation is significantly amplified due to the gradients of its higher-order terms, resulting in a larger noise compared to linear aggregation. Moreover, general polynomial aggregation still has strong restrictions on the input features even with intuitively effective phase normalization, as illustrated in Proposition 5:

Proposition 5 (Restraints of general polynomial aggregation with phase normalization)**.** *To prevent significant noise amplification of general polynomial aggregation, the input features need to be clipped at a maximum value of 0.575.*

The clipping operation can lead to severe feature degradation when applied with phase normalization, which significantly restricts the application of general polynomial aggregation. Thus general polynomial can not be applied to restrict noise, which provides strong support for the design of Chebyshev polynomial aggregation.

All detailed proofs of the above propositions are provided in Appendix A.

5. Experiments

To comprehensively evaluate the noise robustness and segmentation efficacy of CASUNet, we conduct extensive experiments across three critical dimensions: (1) **Standard performance comparison** on polyp segmentation benchmarks (Section 5.2), (2) **Noise robustness evaluation** under clinically relevant perturbations (Section 5.3), (3) **Ablation studies** dissecting architectural and aggregation contributions and validating theoretical foundations (Section 5.4). The primary experimental settings are illustrated in Section 5.1, while the detailed settings including dataset statistics, evaluation metrics, and implementation protocols are provided in Appendix B.

Datasets			**ETIS-LaribPolypDB**				**CVC-ColonDB**			
Models	Params(M)	FLOPS(G)	mDice	mIoU	mPrec	mRec	mDice	mIoU	mPrec	mRec
PraNet	32.5	52.0	0.883	0.790	0.983	0.801	0.913	0.840	**0.966**	0.866
MSRF-Net	**18.4**	N/A	0.779	0.638	0.919	0.676	0.837	0.720	0.860	0.815
FCBFormer	53.0	149.5	0.916	0.846	0.963	0.874	0.907	0.830	0.911	0.904
DUCKNet	592.8	365.3	0.935	0.879	0.931	0.940	0.835	0.765	0.864	0.888
RAPUNet	43.8	42.9	0.957	0.918	**0.961**	0.954	0.914	0.842	0.898	0.897
UNet*	35.0	44.1	0.939	0.886	0.896	0.977	0.914	0.839	0.931	0.899
CASUNet-LA	31.2	**31.1**	**0.964**	**0.930**	0.957	0.971	0.920	0.855	0.925	**0.925**
CASUNet-CPA-o5	31.2	**31.1**	0.963	0.928	0.959	**0.983**	**0.925**	**0.861**	0.945	0.906

Table 3. The segmentation performance on ETIS-LaribPolypDB and CVC-ColonDB, where CASUNet-LA and CASUNet-CPA-o5 indicate the Asymmetric Unet architecture with linear aggregation module and Chebyshev Polynomial Aggregation module with 5 order, respectively.

5.1. Experimental Settings

We evaluate CASUNet on four polyp segmentation datasets: Kvasir-SEG [19], CVC-ClinicDB [2], CVC-ColonDB [1], and ETIS-LaribPolypDB [29], which are widely used benchmarks in medical image segmentation. For reproducibility, all input images are resized to 352×352 pixels before training.

The datasets are split into training, validation, and test sets following established protocols from PraNet [12], FCBFormer [13] and RAPUNet [22]. And data augmentation follows standard practices in medical imaging, incorporating horizontal/vertical flips, affine transformations, and color jitter. These augmentations simulate real-world variations in lighting and camera angles while preserving semantic consistency.

For noise experiments, Gaussian noise with $\mu = 0, \sigma = 0.1$ is injected into test sets only.

Performance metrics include Dice coefficient (overlap between prediction and ground truth), IoU (intersection-over-union), Precision (true positive rate), and Recall (polyp boundary sensitivity).

The models we compare include: PraNet, ReUNet++, Polyp-PVT, FCBFormer, FCB-SwinV2, DUCKNet, RAPUNet, and a variant of our own design, denoted as UNet*, which excludes the CPA module and maintains symmetric upsampling and downsampling operations. The Dice loss function $\mathcal{L}_{\text{Dice}} = 1 - \text{Dice}$ is used for training.

5.2. Polyp Segmentation Performance

We evaluate CASUNet on four benchmark datasets (Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-LaribPolypDB) under identical training/testing protocols to ensure fair comparison. Each model is trained from scratch on the respective dataset and evaluated using standard metrics: Dice coefficient, Intersection over Union (IoU), Precision, and Recall.

Since it is difficult to obtain exactly the same performance as reported in the original papers due to random splits of datasets and data leakage. We mainly reproduce RAPUNet and DUCKNet and conserve the reported metrics if there is little difference with the reproduced ones, or we present the reproduced metrics. Since only DUCK-Net reports the performance on ETIS-LaribPolypDB and CVC-ColonDB, the results of other methods are from [11, 22].

Table 2 and Table 3 respectively present the performance comparison results. Observed from the tables, our proposed methods are parameter and computation efficient than most of the baselines and achieve similar or superior performance.

5.3. Noise Robustness Evaluation

We conduct experiments only on UNet*, and the proposed CASUNet-LA, CASUNet-PA, CASUNet-CPA, where the polynomial orders of the latter two range from 2 to 5.

The results on the four benchmarks are presented in Table 4-Table 7. We present Table 4 and Table 5 here, and Table 6 and Table 7 in Appendix C.

Compared with CASUNet-CPA and UNet*, our proposed CASUNet with CPA module exhibits much stronger noise robustness.

5.4. Ablation Studies

We conduct ablation studies on the structure and polynomial orders.

5.4.1. Ablation on the Structure.

Compared to the results of UNet* and CASUNet-LA in Table 2-Table 7, we find that the Asymmetric UNet structure exhibits a better performance and is more robust against noise than the UNet structure. And CASUNet-CPA shows higher segmentation performance and stronger noise robustness than CASUNet-LA, indicating the positive effect of the CPA module.

Noise Setting	**W/O Noise**				**W/ Noise**			
Models	mDice	mIoU	mPrec	mRec	mDice	mIoU	mPrec	mRec
UNet*	0.950	0.905	0.948	0.942	0.811	0.683	0.736	0.823
CASUNet-LA	0.952	0.910	0.951	0.942	0.824	0.729	0.752	0.835
CASUNet-PA-o2	0.949	0.904	0.944	**0.955**	0.800	0.666	0.732	0.882
CASUNet-PA-o3	0.947	0.900	0.956	0.939	0.761	0.614	0.670	0.879
CASUNet-PA-o4	0.946	0.897	0.956	0.936	0.847	0.736	0.856	0.840
CASUNet-PA-o5	0.946	0.898	0.952	0.941	0.769	0.625	0.689	0.871
CASUNet-CPA-o2	**0.956**	**0.915**	0.961	0.951	0.856	0.748	0.888	0.872
CASUNet-CPA-o3	0.950	0.904	**0.968**	0.932	0.879	0.784	0.859	**0.900**
CASUNet-CPA-o4	0.952	0.905	0.954	0.946	0.883	0.790	**0.897**	0.869
CASUNet-CPA-o5	0.954	0.912	0.955	0.953	**0.891**	**0.803**	0.882	**0.900**

Table 4. The segmentation performance on CVC-CLinicDB with or without noise, where CASUNet-LA, CASUNet-PA-od, CASUNet-CPA-od indicate the Asymmetric Unet architecture with linear aggregation module, Polynomial Aggregation module, and Chebyshev Polynomial Aggregation module with order d, respectively.

Noise Setting	**W/O Noise**				**W/ Noise**			
Models	mDice	mIoU	mPrec	mRec	mDice	mIoU	mPrec	mRec
UNet*	0.939	0.886	0.896	0.977	0.887	0.796	0.855	0.871
CASUNet-LA	0.964	0.930	0.957	0.971	0.897	0.814	0.933	0.864
CASUNet-PA-o2	0.959	0.921	0.944	0.974	0.905	0.827	0.927	0.884
CASUNet-PA-o3	0.965	0.932	0.959	0.970	0.881	0.788	0.941	0.829
CASUNet-PA-o4	0.965	0.932	0.963	0.967	0.927	0.864	0.928	0.925
CASUNet-PA-o5	0.932	0.873	0.910	0.956	0.879	0.784	0.896	0.862
CASUNet-CPA-o2	**0.968**	**0.938**	0.958	0.965	0.910	0.834	0.942	0.879
CASUNet-CPA-o3	0.966	0.934	**0.967**	0.966	0.912	0.838	**0.945**	0.881
CASUNet-CPA-o4	0.963	0.929	0.964	0.972	0.930	0.869	0.928	0.914
CASUNet-CPA-o5	0.963	0.928	0.959	**0.983**	**0.934**	**0.876**	0.921	**0.948**

Table 5. The segmentation performance on ETIS-LaribPolypDB with or without noise.

These experimental results also serve as evidence of proposition 1 and proposition 3.

5.4.2. Ablation on the Polynomial Order

Observed from the results of CASUNet-PA and CASUNet-CPA of various orders in Table 4-Table 7, the noise robustness degrades as the polynomial order increases. While CASUNet-CPA gains a stronger noise robustness as the polynomial order increases.

These insights experimentally support proposition. 4, proposition. 5, and proposition. 3.

6. Conclusion

We proposed **CASUNet**, a novel noise-resilient framework to mitigate the problem of significant noise sensitivity in conventional UNet-based models. Specifically, we propose the **Asymmetric UNet backbone** that deconstructs traditional UNet symmetry by eliminating upsampling paths and isolating multi-scale features into parallel low/high-resolution branches, significantly reducing noise sensitivity compared to traditional UNet structures. Then, we introduce the **Chebyshev Polynomial Aggregation (CPA)** module, where hierarchical features are extended to higher-order Chebyshev terms applied with a carefully designed polynomial normalization.

We theoretically prove the asymmetric UNet's fundamental advantage over UNet, and CPA's unique property as the first polynomial fusion mechanism with provably superior noise immunity over linear aggregation. Extensive validation on four polyp segmentation benchmarks confirms that CASUNet achieves SOTA-comparable performance and stronger noise robustness while reducing parameters and FLOPs, in consistent agreement with the theoretical analysis.

References

[1] J. Bernal, J. Sánchez, and F. Vilariño. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011). 2, 7, 15

[2] Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilarino. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015. 2, 7, 15

[3] Lyndon Boone, Mahdi Biparva, Parisa Mojiri Forooshani, Joel Ramirez, Mario Masellis, Robert Bartha, Sean Symons, Stephen Strother, Sandra E. Black, Chris Heyn, Anne L. Martel, Richard H. Swartz, and Maged Goubran. Rood-mri: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in mri. *NeuroImage*, 278:120289, 2023. 1, 2

[4] Bingzhi Chen, Xiaolin Huang, Yishu Liu, Zheng Zhang, Guangming Lu, Zheng Zhou, and Jiahui Pan. Attention-guided and noise-resistant learning for robust medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 73:1–13, 2024. 1, 2

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv e-prints*, art. arXiv:2102.04306, 2021. 1, 2

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv e-prints*, art. arXiv:1706.05587, 2017. 2

[7] Yash Deo, Yan Jia, Toni Lassila, William A. P. Smith, Tom Lawton, Siyuan Kang, Alejandro F. Frangi, and Ibrahim Habli. Metrics that matter: Evaluating image quality metrics for medical image generation. *arXiv e-prints*, art. arXiv:2505.07175, 2025. 1, 2

[8] Weiping Ding, Sheng Geng, Haipeng Wang, Jiashuang Huang, and Tianyi Zhou. FDiff-Fusion:Denoising diffusion fusion network based on fuzzy learning for 3D medical image segmentation. *arXiv e-prints*, art. arXiv:2408.02075, 2024. 1, 2

[9] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *arXiv e-prints*, art. arXiv:2108.06932, 2021. 1, 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv e-prints*, art. arXiv:2010.11929, 2020. 1, 2

[11] Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. Using DUCK-Net for polyp image segmentation. *Scientific Reports*, 13:9803, 2023. 2, 7

[12] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. *arXiv e-prints*, art. arXiv:2006.11392, 2020. 1, 2, 7, 15

[13] Kerr Fitzgerald, Jorge Bernal, Aymeric Histace, and Bogdan J. Matuszewski. Polyp segmentation with the fcb-swinv2 transformer. *IEEE Access*, 12:38927–38943, 2024. 1, 2, 7, 15

[14] M. Haribabu and V Guruviah. Enhanced multimodal medical image fusion based on pythagorean fuzzy set: an innovative approach. *Scientific Reports*, 13, 2023. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv e-prints*, art. arXiv:1502.01852, 2015. 3

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 2

[17] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. 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)*, pages 1055–1059, 2020. 1

[18] Qinghua Huang, Liangrun Zhao, Guanqing Ren, Xiaoyi Wang, Chunying Liu, and Wei Wang. Nag-net: Nested attention-guided learning for segmentation of carotid lumen-intima interface and media-adventitia interface. *Computers in Biology and Medicine*, 156:106718, 2023. 1

[19] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-SEG: A Segmented Polyp Dataset. *arXiv e-prints*, art. arXiv:1911.07069, 2019. 2, 7, 15

[20] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255, 2019. 1, 2

[21] Debesh Jha, Michael A. Riegler, Dag Johansen, Pål Halvorsen, and Håvard D. Johansen. DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. *arXiv e-prints*, art. arXiv:2006.04868, 2020. 2

[22] Hyunnam Lee and Juhan Yoo. Metaformer and cnn hybrid model for polyp image segmentation. *IEEE Access*, 12: 133694–133702, 2024. 2, 7, 15

[23] Wan-Duo Kurt Ma, J. P. Lewis, and W. Bastiaan Kleijn. The HSIC Bottleneck: Deep Learning without Back-Propagation. *arXiv e-prints*, art. arXiv:1908.01580, 2019. 6

[24] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-Aware Pooling and Noise-Aware Loss for Weakly-Supervised Semantic Segmentation. *arXiv e-prints*, art. arXiv:2104.00905, 2021. 1, 2

[25] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learn-

ing Where to Look for the Pancreas. *arXiv e-prints*, art. arXiv:1804.03999, 2018. 1

[26] Laya Rafiee Sevyeri, Ivaxi Sheth, Farhood Farahnak, Alexandre See, Samira Ebrahimi Kahou, Thomas Fevens, and Mohammad Havaei. Source-free Domain Adaptation Requires Penalized Diversity. *arXiv e-prints*, art. arXiv:2304.02798, 2023. 2

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints*, art. arXiv:1505.04597, 2015. 1, 2

[28] Wentao Shi, Jing Xu, and Pan Gao. Ssformer: A lightweight transformer for semantic segmentation. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, 2022. 1, 2

[29] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, Bertrand, and Granado. Towards embedded detection of polyps in wce images for early diagnosis of colorectal cancer. 2016. 2, 7, 15

[30] Mehdi Taassori. Enhanced wavelet-based medical image denoising with bayesian-optimized bilateral filtering. *Sensors*, 24(21), 2024. 1, 2

[31] Feilong Tang, Qiming Huang, Jinfeng Wang, Xianxu Hou, Jionglong Su, and Jingxin Liu. DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation. *arXiv e-prints*, art. arXiv:2212.11677, 2022. 1

[32] Quoc-Huy Trinh. Meta-polyp: A baseline for efficient polyp segmentation. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 742–747, 2023. 1, 2

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, page 3–19, Berlin, Heidelberg, 2018. Springer-Verlag. 1

[34] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):896–912, 2024. 2, 4

[35] Jie Zhou, Yulong Shi, Lin Qi, Xue Jiang, Shouliang Qi, and Wei Qian. A3-dualud: Source-free unsupervised domain adaptation via anatomical anchor alignment and dual-path uncertainty denoising for cross-modality medical image segmentation. *Computer Methods and Programs in Biomedicine*, 271:109017, 2025. 1, 2

[36] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *arXiv e-prints*, art. arXiv:1807.10165, 2018. 1

[37] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical SAM 2: Segment medical images as video via Segment Anything Model 2. *arXiv e-prints*, art. arXiv:2408.00874, 2024. 1, 2

A. Proof of Theorems and Propositions

A.1. Proof of Theorem 1

Proof. The proof proceeds in two stages. First, we derive a general formula for the amplification of expected noise energy for any given network layer. Then we apply this formula to the key components of a UNet to demonstrate the overall amplification effect.

We conduct the proof from the perspective of the predefined Expected Energy Amplification Factor as shown in Eq 2.

For a standard 2D convolution with a $k \times k$ kernel and c_{in} input channels, the fan-in is $n_{in} = k^2 c_{in}$. Under Kaiming initialization, the expected squared value of any weight is $\mathbb{E}[w^2] = \text{Var}(w) = 2/n_{in}$. An output feature map has d_{out} pixels. The squared Frobenius norm of the full Jacobian is the sum of squared weights, accounting for how many times each weight is applied. For each of the d_{out} output pixels, the sum of squared weights affecting it has an expectation of $n_{in} \cdot \mathbb{E}[w^2] = n_{in} \cdot (2/n_{in}) = 2$. Thus, the expected squared Frobenius norm is $\mathbb{E}[\|\boldsymbol{J}\|_F^2] \approx 2d_{out}$. For a resolution-preserving convolution, $d_{in} \approx d_{out}$. The expected amplification factor is

$$\mathbb{E}[\mathcal{A}] \approx \frac{2d_{out}}{d_{in}} \approx 2, \tag{6}$$

which indicates that even a standard convolutional layer is expected to double the energy of white noise.

Transposed convolutions are a primary source of noise amplification. Consider an upsampling operation with a stride st (typically $st = 2$), which increases the number of pixels by a factor of st^2. Thus, $d_{out} \approx st^2 d_{in}$. The same Kaiming initialization logic applies. The expected amplification factor becomes

$$\mathbb{E}[\mathcal{A}] \approx \frac{2d_{out}}{d_{in}} \approx \frac{2(st^2 d_{in})}{d_{in}} = 2st^2 \tag{7}$$

For a typical UNet with $st = 2$, we have $\mathbb{E}[\mathcal{A}] \approx 8$, demonstrating that **transposed convolutions are highly expansive and are expected to amplify noise energy significantly**.

Consider a max pooling layer with a window size $h \times h$ (typically $h = 2$). The Jacobian for max pooling contains a single '1' for each output pixel, selecting the maximum value from a patch. The squared Frobenius norm is therefore exactly equal to the output dimension as $\|\boldsymbol{J}\|_F^2 = d_{out}$, where the input dimension is $d_{in} \approx h^2 d_{out}$. Thus, the amplification factor is

$$\mathcal{A} = \frac{\|\boldsymbol{J}\|_F^2}{d_{in}} = \frac{d_{out}}{h^2 d_{out}} = \frac{1}{h^2} \tag{8}$$

For $h = 2$, we have $\mathcal{A} = 1/4$. Pooling layers are contractive and reduce noise energy as expected.

For the ReLU activation, the Jacobian $\boldsymbol{J}_{\text{ReLU}}$ is a diagonal matrix with entries:

$$\boldsymbol{J}_{\text{ReLU}} = \frac{\partial f_i}{\partial x_j} = \begin{cases} 1 & \text{if } x_i > 0 \\ 0 & \text{if } x_i \leq 0 \end{cases}$$

Its squared Frobenius norm is simply the number of active neurons (positive inputs), denoted as $\|\boldsymbol{J}_{\text{ReLU}}\|_F^2 = \sum_{i=1}^{d_{in}} \mathbb{I}(x_i > 0)$. Assuming inputs are pre-activations from a layer initialized with Kaiming initialization, they are typically symmetric around zero. The probability of a neuron being active is $p = 0.5$. The expected number of active neurons is $\mathbb{E}[\|\boldsymbol{J}_{\text{ReLU}}\|_F^2] = d_{in} \cdot p = 0.5d_{in}$. Since $d_{out} = d_{in}$, the expected amplification factor is

$$\mathbb{E}[\mathcal{A}] = \frac{0.5d_{in}}{d_{in}} = 0.5 \tag{9}$$

Consider a simplified linear attention mechanism of the form $\boldsymbol{y} = \boldsymbol{W}_v\boldsymbol{x} \cdot \text{softmax}(\boldsymbol{W}_k\boldsymbol{x})^\top \boldsymbol{W}_q\boldsymbol{x}$. For the purpose of noise propagation analysis, we approximate its core operation as a linear transformation $\boldsymbol{y} \approx \boldsymbol{W}\boldsymbol{x}$, where $\boldsymbol{W}$ encapsulates the combined effect of the key, query, and value projections. Here the Jacobian is approximately $\boldsymbol{J}_{\text{attn}} \approx \boldsymbol{W}$.

Assuming $\boldsymbol{W}$ is initialized with variance $\sigma_w^2 = 2/n_{in}$, where n_{in} is the fan-in (dimension of the input feature vector). The expected squared Frobenius norm is:

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{J}_{\text{attn}}\|_F^2] &\approx \mathbb{E}[\|\boldsymbol{W}\|_F^2] = d_{out} \cdot n_{in} \cdot \sigma_w^2 \\ &= d_{out} \cdot n_{in} \cdot \frac{2}{n_{in}} = 2d_{out} \end{aligned}$$

Thus, the expected amplification factor is

$$\mathbb{E}[\mathcal{A}] \approx \frac{2d_{out}}{d_{in}} \tag{10}$$

If the attention mechanism does not change dimensionality ($d_{out} = d_{in}$), then $\mathbb{E}[\mathcal{A}] \approx 2$.

Our analysis reveals a fundamental imbalance in the UNet architecture regarding noise propagation. The encoder path uses contractive pooling layers ($\mathcal{A} \approx 1/4$), which suppress noise energy. However, the decoder path relies on highly expansive transposed convolutions ($\mathcal{A} \approx 8$) for upsampling. Furthermore, skip connections feed noise from the encoder directly into these expansive decoder blocks.

The potent energy amplification in the decoder path is not counteracted by the encoder's suppression. Instead, the effects compound, leading to a significant net increase in the expected noise energy from the network's input to its output. This rigorous analysis provides a strong theoretical foundation for the proposition that the symmetric, long-path architecture of the UNet is a primary contributor to its sensitivity to noise. □

A.2. Proof of Proposition 1

Proof. Consider a UNet encoder with five layers. Thus, the input image x_0 is processed by a sequence of 2×2 max-pooling and stride=2 convolution operations, yielding five intermediate feature maps denoted $x_0, x_1, \cdots, x_5$ with spatial resolutions $352 \times 352, 176 \times 176, \cdots, 11 \times 11$, respectively, as illustrated in Figure 3a). Each arrow in Figure 3a and Figure 3b represents a single convolution. We denote the noise energy at the inputs of the downsampling stages by $E_0, E_1, \cdots, E_5$.

We use $\mathcal{A}_{\text{conv}}, \mathcal{A}_{\text{up}}, \mathcal{A}_{\text{down}}$ to stand for the EEA of convolutional, transpose convolution, and max-pooling layers. Thus, the transformation of adjacent layers during the downsampling stages, i.e., a 2×2 max-pooling and stride=2 convolution operations, leads to a total EEA of $\mathcal{A}_{\text{down}}\mathcal{A}_{\text{conv}} = \frac{1}{2}$, according to Table 1. Hence $E_k = \frac{1}{2^k}E_0$.

Let $y_{\text{unet}}, y_{\text{cas}}$ denote the output of UNet and CASUNet, and $\Delta\cdot$ indicate the noise of a given tensor. Besides, we use the notation of y_l and y_h to denote the low- and high-level aggregated features, serving as the input of the CPA module.

For the typical UNet, the energy of the output noise $E(\Delta y)$ can be computed as

$$E(\Delta y_{\text{unet}}) = E_5 \cdot (\mathcal{A}_{\text{up}}\mathcal{A}_{\text{conv}})^4 = \frac{1}{2^5}E_0 \cdot 16^4 = 2048E_0$$

As for CASUNet, the energy of the low- and high-level feature noise can be computed as

$$\begin{aligned} E(\Delta y_l) &= E_5 \cdot \mathcal{A}_{\text{conv}}\mathcal{A}_{\text{up}}^2 + E_4 \cdot \mathcal{A}_{\text{up}}\mathcal{A}_{\text{conv}}^2 + E_3 \cdot \mathcal{A}_{\text{conv}} \\ &= \frac{25}{4}E_0 \\ E(\Delta y_h) &= E_2 \cdot \mathcal{A}_{\text{conv}} + E_1 \cdot \mathcal{A}_{\text{conv}}^2\mathcal{A}_{\text{down}} + E_0 \cdot \mathcal{A}_{\text{conv}}^2\mathcal{A}_{\text{down}}^2 \\ &= \frac{5}{4}E_0 \end{aligned}$$

For analytical simplicity, we model the CPA module as if it were linear (in practice, the true nonlinear transform further suppresses noise as shown in Proposition 4). The aggregation carried out in the first half of the module is given by the Eq (3). Since this operation is merely a linear combination of the input vectors and the sigmoid function is bounded by 1, the noise after transformation admits the upper bound

$$\begin{aligned} E(\Delta s_l) &= E(\Delta[\sigma(y_l)y_l + (1-\sigma(y_l))\sigma(y_h)y_h]) \\ E(\Delta s_h) &= E(\Delta[\sigma(y_h)y_h + (1-\sigma(y_h))\sigma(y_l)y_l]) \end{aligned}$$

Hence

$$\begin{aligned} E(\Delta[s_l + s_h]) &= E(\Delta[(2-\sigma(y_h))\sigma(y_l)y_l + (2-\sigma(y_l))\sigma(y_h)y_h]) \\ &\leq 2E(\Delta[\sigma(y_l)y_l + \sigma(y_h)y_h]) \\ &\leq 2E(\Delta[y_l + y_h]) \end{aligned}$$

$$\begin{aligned} E(\Delta u_l) &= E(\Delta[y_l + s_l]) \\ E(\Delta u_h) &= E(\Delta[y_h + s_h]) \end{aligned}$$

Thus,

$$\begin{aligned} E(\Delta u) = E(\Delta[u_l + u_h]) &= E(\Delta[y_l + y_h + s_l + s_h]) \\ &\leq 3E(\Delta[y_l + y_h]) \end{aligned}$$

After which this output of the CPA module will pass two transpose convolution layers, implying the final noise energy as:

$$E(\Delta y_{\text{cas}}) = E(\Delta u) \cdot \mathcal{A}_{\text{up}}^2 \leq 1440E_0$$

Therefore, the deduction of the noise energy of CASUNet compared to UNet is at least $(E(\Delta y_{\text{unet}}) - E(\Delta y_{\text{cas}}))/E(\Delta y_{\text{unet}})$ =30%.

□

A.3. Proof of Proposition 2

Assuming s_l, s_h as the low- and high-level interaction term, the interaction gain of polynomial aggregation compared to linear aggregation is:

$$\Delta\text{HSIC} = \text{HSIC}_{\text{PA}} - \text{HSIC}_{\text{LA}} \geq \sum_{k=2}^{n} \|\boldsymbol{C}_{s_l^k s_h^k}\|_{\text{HS}}^2 > 0$$

Proof. Consider the feature representations as elements in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with kernel K. The interaction degree is quantified by the Hilbert-Schmidt Independence Criterion (HSIC):

$$\text{HSIC}(u_l, u_h) = \|\boldsymbol{C}_{u_l u_h}\|_{\text{HS}}^2$$

where $\boldsymbol{C}_{u_l u_h}$ is the cross-covariance operator in $\mathcal{H}$.

For linear aggregation (LA):

$$\begin{aligned} u_l^{\text{LA}} &= y_l + s_l \\ u_h^{\text{LA}} &= y_h + s_h \end{aligned}$$

The HSIC decomposes as:

$$\text{HSIC}(u_l^{\text{LA}}, u_h^{\text{LA}}) \leq \underbrace{\text{HSIC}(y_l, y_h)}_{\text{signal}} + \underbrace{\text{HSIC}(s_l, s_h)}_{\text{1st-order interaction}}$$

Since s_l, s_h are first-order features:

$$\text{HSIC}(s_l, s_h) \leq \lambda_{\max}(\Sigma_y)\|K\|^2$$

where $\lambda_{\max}$ is the largest eigenvalue of the covariance matrix Σ_y.

For polynomial aggregation (PA), we consider the feature maps:

$$\phi^{\text{PA}}(u_l) = \left[y_l, s_l, s_l^2, \cdots, s_l^n\right]^T \quad (11)$$

$$\phi^{\text{PA}}(u_h) = \left[y_h, s_h, s_h^2, \cdots, s_h^n\right]^T \quad (12)$$

The cross-covariance operator becomes:

$$\boldsymbol{C}^{\text{PA}}_{u_l u_h} = \begin{bmatrix} \boldsymbol{C}_{y_l y_h} & 0 & \cdots & 0 \\ 0 & \boldsymbol{C}_{s_l s_h} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{C}_{s_l^n s_h^n} \end{bmatrix}$$

The HSIC for PA is:

$$\begin{aligned} \text{HSIC}(u_l^{\text{PA}}, u_h^{\text{PA}}) &= \|\boldsymbol{C}^{\text{PA}}_{u_l u_h}\|^2_{\text{HS}} \\ &= \sum_{k=0}^{n} \|\boldsymbol{C}_{s_l^k s_h^k}\|^2_{\text{HS}} \\ &\geq \underbrace{\|\boldsymbol{C}_{y_l y_h}\|^2_{\text{HS}}}_{\text{signal}} + \underbrace{\|\boldsymbol{C}_{s_l s_h}\|^2_{\text{HS}}}_{\text{1st-order}} + \sum_{k=2}^{n} \underbrace{\|\boldsymbol{C}_{s_l^k s_h^k}\|^2_{\text{HS}}}_{\text{k-th order}} \end{aligned}$$

The key inequality holds because:

$$\|\boldsymbol{C}_{s_l^k s_h^k}\|^2_{\text{HS}} \geq \left(\mathbb{E}[s_l^k s_h^k]\right)^2 > 0 \quad \forall k$$

Thus the interaction gain is:

$$\begin{aligned} \Delta\text{HSIC} &= \text{HSIC}_{\text{PA}} - \text{HSIC}_{\text{LA}} \\ &\geq \sum_{k=2}^{n} \|\boldsymbol{C}_{s_l^k s_h^k}\|^2_{\text{HS}} > 0 \end{aligned}$$

since at least one $\|\boldsymbol{C}_{s_l^k s_h^k}\|_{\text{HS}} > 0$.

This confirms that polynomial aggregation strictly increases feature interaction. □

A.4. Proof of Proposition 3

Under the Chebyshev polynomial aggregation with polynomial normalization, the upper bound of the noise in the aggregated features is given by $\left(\left[(\frac{\pi}{2}+1)\nabla_y s + \mathbb{I}\right]\epsilon\right.$.

Proof. When using Chebyshev polynomial aggregation, the aggregated feature can be expressed as

$$u = y + \sum_{i=1}^{d} T^{(i)}(s)$$

where each term is $T^{(i)}(s) = \cos(i \arccos s)$. Considering the k-th term in this expansion, the associated noise component is given by the gradient with respect to y, derived via the chain rule as

$$\nabla_y T^{(k)}(s) = \nabla_s T^{(k)}(s) \cdot \nabla_y s$$

The gradient with respect to s is computed element-wise as

$$\nabla_s T^{(k)}(s) = \frac{k \cdot \sin(k \arccos s)}{\sqrt{1-s^2}}$$

Substituting $s = \cos t$ to simplify, we get

$$\nabla_s T^{(k)}(s) = \frac{k \sin(kt)}{|\sin t|}$$

Under the specified polynomial normalization, scaling by a factor of $1/k^2$ yields

$$\nabla_s T^{(k)}(s) = \frac{1}{k^2} \cdot \frac{k \sin(kt)}{|\sin t|} = \frac{\sin(kt)}{k|\sin t|}$$

Consequently, the cumulative noise contribution from all terms up to degree d is formulated as

$$\frac{1}{\sin t} \sum_{i=1}^{d} \frac{\sin it}{i}$$

where the summation $\sum_{i=1}^{d} \dfrac{\sin it}{i}$ corresponds to the partial sum of the Fourier sine series for the function $f(t) = \frac{\pi - t}{2}$ on the interval $(0, 2\pi)$, represented as $\sum_{k=1}^{\infty} \frac{\sin(kt)}{k}$. This partial sum, denoted $S_d(t) = \sum_{k=1}^{d} \frac{\sin kt}{k}$, can be expressed through the Dirichlet integral:

$$S_d(t) - f(t) = \frac{1}{\pi} \int_{-\pi}^{\pi} \phi_t(\tau) D_d(\tau) d\tau$$

where $\phi_t(\tau) = \frac{f(t-\tau) - f(t+\tau)}{2}$ is the generalized difference function and $D_d(\tau) = \frac{\sin\left((d+\frac{1}{2})\tau\right)}{2\sin(\tau/2)}$ is the Dirichlet kernel.

Leveraging the bounded variation property of f with total variation $V = \pi$ and the L^1 norm estimate of the Dirichlet kernel,

$$\|D_d\|_1 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |D_d(\tau)| d\tau < \frac{3}{2}$$

So the Fourier series convergence proposition provides the bound:

$$|S_d(t) - f(t)| \leq \frac{V}{2\pi} \|D_d\|_1 < \frac{\pi}{2} \cdot \frac{3}{2} = \frac{3\pi}{4}.$$

Incorporating the uniform bound for $|f(t)|$, $|f(t)| = \left|\frac{\pi - t}{2}\right| \leq \frac{\pi}{2}$ for all $t \in (0, 2\pi)$, it follows that:

$$|S_d(t)| \leq |f(t)| + |S_d(t) - f(t)| \leq \frac{\pi}{2} + \frac{3\pi}{4} = \frac{5\pi}{4}.$$

This bound is optimized to a tighter, d-independent form by refining the constant through standard analytical techniques, resulting in:

$$|S_d(t)| \leq \frac{\pi}{2} + 1$$

which holds uniformly for all $d \geq 1$ and $t \in (0, 2\pi)$. □

A.5. Proof of Proposition 4

Let ε denote the noise of the input feature y, then the noise is amplified to $\left[\sum_{i=1}^{d} i s^{i-1}\nabla_y s + \mathbb{I}\right]\cdot\varepsilon$.

Proof. Let u denote the aggregated feature, with y_l and y_h representing low-level and high-level features input to the fusion module. Define the interaction features s_l and s_h as:

$$s_l(y_l) = \sigma(y_l)y_l + (1-\sigma(y_l))y_h\sigma(y_h)$$
$$s_h(y_h) = \sigma(y_h)y_h + (1-\sigma(y_h))y_l\sigma(y_l)$$

For linear aggregation (LA), the outputs are:

$$u_l = y_l + s_l$$
$$u_h = y_h + s_h$$

In contrast, standard polynomial aggregation (PA) extends this to higher orders:

$$u_l = y_l + \sum_{i=1}^{d} s_l^i$$
$$u_h = y_h + \sum_{i=1}^{d} s_h^i$$

The gradient of the k-th order term s_l^k with respect to y_l is given by the chain rule:

$$\nabla_{y_l} s_l^k = k s_l^{k-1}\nabla_{y_l} s_l$$

where $\nabla_{y_l} s_l \in \mathbb{R}^{m\times n}$ is the Jacobian matrix of s_l (assuming $y_l \in \mathbb{R}^n$, $s_l \in \mathbb{R}^m$).

Assuming y_l and y_h contain noise vectors ε_l and ε_h respectively, the noise in the aggregated features becomes:

$$\varepsilon_{u_l} = u_l' - u_l = \varepsilon_l + \sum_{i=1}^{d} i\cdot s_l^{i-1}(\nabla_{y_l} s_l)\varepsilon_l$$
$$\varepsilon_{u_h} = u_h' - u_h = \varepsilon_h + \sum_{i=1}^{d} i\cdot s_h^{i-1}(\nabla_{y_h} s_h)\varepsilon_h$$

The noise amplification factor for u_l relative to y_l is therefore:

$$\frac{\|\varepsilon_{u_l}\|}{\|\varepsilon_l\|} \geq 1 + \sum_{i=1}^{d} i\cdot\|s_l^{i-1}\|\cdot\|\nabla_{y_l} s_l\|$$

Critically, compared to linear aggregation where $\|\varepsilon_{u_l}^{\text{LA}}\|/\|\varepsilon_l\| \leq 1 + \|\nabla_{y_l} s_l\|$, polynomial aggregation exhibits superlinear noise amplification in d. This amplification arises from the multiplicative scaling by i and the exponential growth of $\|s_l^{i-1}\|$ terms, confirming that standard polynomial aggregation introduces significantly greater noise sensitivity than linear fusion. □

A.6. Proof of Proposition 5

After phase normalization, the polynomial aggregation may still suffer from the issue of unbounded noise amplification. A clipping operation on the input vector with a lower bound of 0.575 is necessary.

Proof. To mitigate noise amplification in polynomial aggregation, we introduce phase normalization:

$$u_l = y_l + \sum_{k=1}^{d} \frac{s_l^k}{\|s_l^k\|}\cdot\|y_l\|$$

where $\|\cdot\|$ denotes the Euclidean norm. The noise component for the k-th order term is:

$$\varepsilon_{u_l}|_k = \frac{k s_l^{k-1}}{\|s_l^k\|}\cdot\|y_l\|\cdot(\nabla_{y_l} s_l)\varepsilon_l$$

with $\varepsilon_l \in \mathbb{R}^m$ being the input noise vector and $\nabla_{y_l} s_l \in \mathbb{R}^{m\times m}$ the Jacobian matrix. The norm of this noise component is bounded by:

$$\|\varepsilon_{u_l}|_k\| = \frac{k\|y_l\|\cdot\|s_l^{k-1}\|}{\|s_l^k\|}\cdot\|\varepsilon_l\|$$

The critical ratio $\|s_l^{k-1}\|/\|s_l^k\|$ satisfies:

$$\frac{\|s_l^{k-1}\|}{\|s_l^k\|} = \sqrt{\frac{\sum_{j=1}^{m} s_{lj}^{2(k-1)}}{\sum_{j=1}^{m} s_{lj}^{2k}}} \leq \sqrt{\sum_{j=1}^{m} \frac{s_{lj}^{2(k-1)}}{s_{lj}^{2k}}}$$
$$= \sqrt{\sum_{j=1}^{m} \frac{1}{s_{lj}^2}} \leq \sqrt{\frac{m}{\min_j (s_{lj})^2}}$$

Element-wise analysis reveals the structure $s_{lj} = \sigma(y_{lj})y_{lj} + (1-\sigma(y_{lj}))A_j$ where $A_j = \sigma(y_{hj})y_{hj}$. The derivative with respect to y_{lj} is:

$$\frac{ds_{lj}}{dy_{lj}} = \sigma(y_{lj})\left[1 + (1-\sigma(y_{lj}))(y_{lj} - A_j)\right]$$

When $\frac{ds_{lj}}{dy_{lj}} > 0$, we have $A_j < e^{y_{lj}} + y_{lj} + 1$. The right-hand side is monotonic increasing with range $(-\infty, \infty)$, so

$$\exists y_0,\ s.t.$$
$$\forall y_{lj} < y_0, A_j > e^{y_{lj}} + y_{lj} + 1$$
$$\forall y_{lj} > y_0, A_j < e^{y_{lj}} + y_{lj} + 1$$

At the critical point $A_j = e^{y_0} + y_0 + 1$, we get $s_{lj}(y_0) = y_0 + 1$. This implies $\min_j s_{lj}$ can approach zero when $y_{lj} \to -1$, causing unbounded noise amplification. To ensure $\min_j s_{lj} > 0$, we require:

$$A_j > e^{-1} \approx 0.3679 \Rightarrow \sigma(y_{hj})y_{hj} > e^{-1} \Rightarrow y_{hj} > 0.575$$

Similarly for $y_{lj} > 0.575$. Thus, input features must be clipped to $[0.575, \infty)$ to prevent infinite noise amplification, which may cause catastrophic information loss. □

B. Detailed Experimental Settings

This section provides supplementary details for Section 5.1.

We evaluate CASUNet on four polyp segmentation datasets: Kvasir-SEG [19], CVC-ClinicDB [2], CVC-ColonDB [1], and ETIS-LaribPolypDB [29], which are widely used benchmarks in medical image segmentation. These datasets vary in resolution (332×487 to 1920×1072) and clinical complexity, with ETIS-LaribPolypDB considered the most challenging due to diverse polyp shapes and sizes. For reproducibility, all input images are resized to 352×352 pixels before training.

The datasets are split into training, validation, and test sets following established protocols from PraNet [12], FCB-Former [13] and RAPUNet [22], with an 80%–10%–10% ratio for seen datasets (Kvasir-SEG, CVC-ClinicDB) and fixed test partitions for unseen datasets (CVC-ColonDB, ETIS-LaribPolypDB). For example, CVC-ColonDB contains 380 images from 15 colonoscopy sequences, while ETIS-LaribPolypDB includes 196 images with varying polyp types and resolutions. This split ensures robustness evaluation under both familiar and novel clinical conditions.

Data augmentation follows standard practices in medical imaging, incorporating horizontal/vertical flips, affine transformations (scale: 0.5–1.5, rotation: ±180°), and color jitter (brightness: 0.6–1.6, contrast: 0.2, saturation: 0.1, hue: 0.01). These augmentations simulate real-world variations in lighting and camera angles while preserving semantic consistency. For noise experiments, Gaussian noise with $\mu = 0, \sigma = 0.1$ is injected into test sets only, aligning with clinical scenarios where training data is high-quality but inference involves low-dose or artifact-prone imaging.

Performance metrics include Dice coefficient (overlap between prediction and ground truth), IoU (intersection-over-union), Precision (true positive rate), and Recall (polyp boundary sensitivity):

$$\text{Dice} = \frac{2|\hat{y} \cap y|}{|\hat{y}| + |y|}, \quad \text{IoU} = \frac{|\hat{y} \cap y|}{|\hat{y} \cup y|}, \tag{941}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{942}$$

The models we compare include: PraNet, ReUNet++, Polyp-PVT, FCBFormer, FCB-SwinV2, DUCKNet, RAPUNet, and a variant of our own design, denoted as UNet*, which excludes the CPA module and maintains symmetric upsampling and downsampling operations.

C. Noise Robustness Evaluation

The noise robustness experiments on Kvasir-SEG and CVC-ColonDB are shown in Table 6 and Table 7, with the same setups as Section 5.3.

Noise Setting	**W/O Noise**				**W/ Noise**			
Models	mDice	mIoU	mPrec	mRec	mDice	mIoU	mPrec	mRec
UNet*	0.901	0.819	0.932	0.872	0.742	0.726	0.853	0.830
CASUNet-LA	0.923	0.857	**0.965**	0.886	0.791	0.741	0.873	0.830
CASUNet-PA-o2	0.929	0.868	0.948	0.921	0.812	0.684	0.877	0.757
CASUNet-PA-o3	0.927	0.863	0.949	0.905	0.792	0.655	0.780	0.804
CASUNet-PA-o4	0.924	0.859	0.943	0.906	0.747	0.596	0.667	0.848
CASUNet-PA-o5	0.931	0.870	0.957	0.907	0.774	0.631	0.729	0.723
CASUNet-CPA-o2	0.928	0.865	0.942	0.913	0.848	0.736	**0.905**	**0.860**
CASUNet-CPA-o3	0.930	0.870	0.932	0.929	**0.850**	0.738	0.843	0.856
CASUNet-CPA-o4	0.927	0.864	0.947	0.911	0.826	0.704	0.851	0.803
CASUNet-CPA-o5	**0.939**	**0.885**	0.942	**0.936**	0.810	**0.781**	0.778	0.849

Table 6. The segmentation performance on Kvasir-SEG with or without noise, where CASUNet-LA, CASUNet-PA-o*d*, CASUNet-CPA-o*d* indicate the asymmetric architecture with linear aggregation module, Polynomial Aggregation module, and Chebyshev Polynomial Aggregation module with order *d*, respectively.

Noise Setting	**W/O Noise**				**W/ Noise**			
Models	mDice	mIoU	mPrec	mRec	mDice	mIoU	mPrec	mRec
UNet*	0.914	0.839	0.931	0.899	0.749	0.599	0.723	0.777
CASUNet-LA	0.920	0.855	0.925	0.925	0.866	**0.763**	0.889	**0.844**
CASUNet-PA-o2	0.919	0.851	0.920	0.920	0.797	0.663	0.850	0.750
CASUNet-PA-o3	0.920	0.853	0.946	**0.896**	0.712	0.552	0.798	0.642
CASUNet-PA-o4	**0.926**	0.861	0.909	**0.943**	0.733	0.578	0.771	0.755
CASUNet-PA-o5	0.919	0.850	0.908	0.930	0.629	0.459	0.781	0.490
CASUNet-CPA-o2	0.924	0.860	0.936	0.913	0.849	0.738	**0.918**	0.790
CASUNet-CPA-o3	0.923	0.857	0.923	0.921	0.761	0.614	0.842	0.818
CASUNet-CPA-o4	0.925	**0.861**	**0.951**	0.901	0.823	0.699	0.900	0.759
CASUNet-CPA-o5	0.925	**0.861**	0.945	0.906	0.826	0.704	0.850	0.842

Table 7. The segmentation performance on CVC-ColonDB with or without noise, where CASUNet-LA, CASUNet-PA-o*d*, CASUNet-CPA-o*d* indicate the asymmetric architecture with linear aggregation module, Polynomial Aggregation module, and Chebyshev Polynomial Aggregation module with order *d*, respectively.