Adaptive Frequency Attention: Frequency-domain Context Modeling for Medical Image Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

U-Net and its variants have become the cornerstone of medical image segmentation by effectively leveraging semantic context to distinguish regions with similar local features. However, convolution-based modules are inherently limited in modeling long-range contextual dependencies. To address this, recent works incorporate Transformer modules to capture global context via self-attention mechanisms. In this work, we explore an alternative direction based on frequency-domain learning, where global dependencies can be captured implicitly and efficiently. We propose Adaptive Frequency Attention (AFA), a lightweight plug-and-play module for U-Net-based architectures that adaptively captures informative frequency components to enhance contextual representations. The AFA module is integrated into the encoder with a residual connection, enabling simultaneous modeling of global context and local details. By applying the Discrete Fourier Transform, spatial information is decomposed into multi-scale frequency components, allowing AFA to adaptively emphasize discriminative spectral responses with a learnable mask. To encourage diverse frequency representations, feature channels are partitioned into groups with an orthogonality regularization to promote complementary spectral focus. In addition, a spectral preservation constraint dynamically regulates suppression strength to prevent the loss of critical frequency information. Extensive experiments on multiple medical image segmentation benchmarks demonstrate that AFA consistently improves U-Net variants and achieves state-of-the-art performance.

1 Introduction

Medical image segmentation aims to partition medical images into meaningful regions corresponding to lesional areas, enabling precise quantitative analysis and supporting downstream clinical tasks such as diagnosis and treatment planning. With the rapid advancement of deep learning, segmentation performance has steadily improved. U-Net [1] and its variants [2, 3, 4, 5, 6, 7, 8, 9] have become the dominant frameworks in medical image segmentation owing to their elegant encoder-decoder architecture, which effectively integrates semantic context with fine-grained details. The encoder progressively downsamples the input to expand the receptive field and capture high-level semantic contextual representations of regions of interest. Then, the decoder gradually upsamples these features to restore spatial resolution and precise localization. This synergy between semantic context and fine-grained detail enables U-Net to outperform traditional fully convolutional networks, particularly in resolving ambiguities in regions with similar local appearances [2].

Although convolutional neural network (CNN)-based encoder-decoder architectures have become the dominant framework for medical image segmentation, their inherently local receptive fields make them more effective at capturing local patterns than modeling long-range contextual dependencies. In contrast, Transformers [10, 11] employ self-attention mechanisms to establish direct interactions between distant spatial regions, enabling more effective global context modeling. Motivated by these

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

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

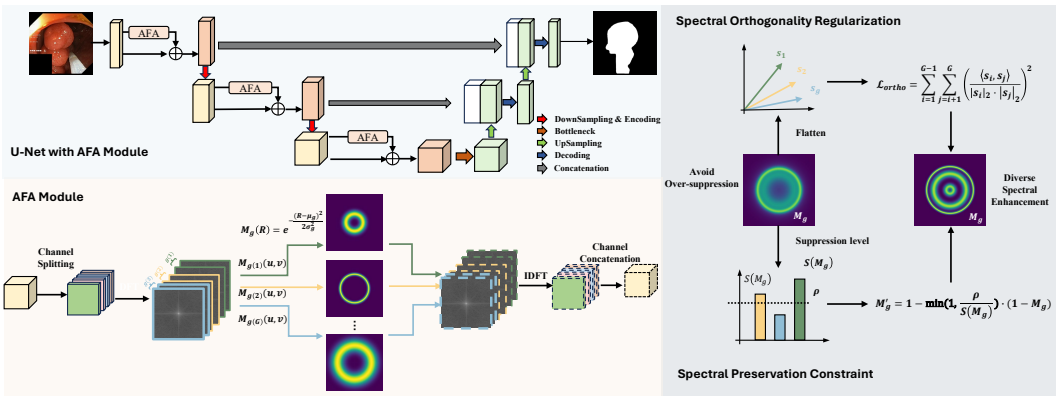


Figure 1: Overview of the proposed Adaptive Frequency Attention (AFA) module integrated into U-Net architecture.

complementary properties, recent studies have explored incorporating Transformers to enhance long-range contextual dependency modeling. For example, TransUNet integrates a Transformer encoder into the U-Net framework to capture global context while preserving fine-grained details [12, 9]. Swin-Unet adopts a pure Transformer backbone within a U-shaped architecture to improve multi-organ segmentation performance [13]. nnWNet revisits convolution-attention fusion under a standardized evaluation protocol, demonstrating the effectiveness of hybrid CNN-Transformer architectures that combine the local feature extraction capability of CNNs with the global dependency modeling ability of Transformers [6].

Although Transformer layers effectively capture long-range dependencies and global context, they typically incur higher computational cost than CNN layers. In contrast, frequency-domain representations provide an efficient alternative by modeling global dependencies implicitly [14]. Motivated by this, we explore a frequency-domain approach for efficient global context modeling in medical image segmentation. To this end, we propose Adaptive Frequency Attention (AFA), a lightweight plug-and-play module that adaptively captures informative frequency components to enhance contextual representations. As shown in Fig. 1, the AFA module is integrated into the encoder of the U-Net structure with a residual connection, enabling simultaneous modeling of long-range dependencies and local details. By applying the Discrete Fourier Transform, spatial information is decomposed into multi-scale frequency components, where AFA employs a ring-shaped Gaussian mask parameterized by learnable center frequency and bandwidth to selectively emphasize discriminative spectral responses. To encourage diverse frequency representations, feature channels are partitioned into groups, and an orthogonality-based regularization is introduced to promote complementary spectral focus. In addition, a spectral preservation constraint dynamically regulates suppression strength to prevent over-attenuation of critical frequency information during training. Designed as a plug-and-play module, AFA can be seamlessly integrated into U-Net-based architectures. Extensive experiments on multiple medical image segmentation benchmarks demonstrate that AFA consistently improves various U-Net variants and achieves state-of-the-art performance.

Our contributions are as follows:

1. We propose Adaptive Frequency Attention (AFA), a lightweight plug-and-play module for U-Net that implicitly and efficiently models both long-range dependencies and local details to improve segmentation accuracy.

2. By applying the Discrete Fourier Transform, AFA learns a mask to emphasize discriminative spectral responses, while promoting diverse frequency representations through channel grouping with an orthogonality-based regularization.

3. Extensive experiments on multiple benchmarks demonstrate that AFA consistently improves U-Net variants, achieving superior localization of lesional regions and state-of-the-art segmentation performance.

2 Related work

2.1 U-Net Architectures

U-Net [1] introduces the encoder-decoder architecture that has become the foundation of modern medical segmentation, effectively combining context aggregation in the encoder with precise spatial localization. Since then, numerous variants have refined this blueprint. Inspired by residual learning [15], residual and densely connected designs (e.g., ResUNet [9] and H-DenseUNet [3]) stabilize optimization and improve feature reuse. To tighten multi-scale interaction and bridge the semantic gap across stages, UNet++ [5] and UNet 3+ [8] redesign skip paths and aggregation. Orthogonally, Attention U-Net [4] injects attention gates to highlight task-relevant regions. Rather than adding new architectural blocks, nnU-Net [2, 16] systematizes the pipeline and remains a robust benchmark.

2.2 Frequency-Domain Learning

Leveraging frequency information has shown strong effectiveness across diverse vision tasks and continues to attract increasing attention [17, 18, 19, 20]. FCA-Net demonstrates that global average pooling is a special case of the Discrete Cosine Transform (DCT), and extends channel attention into the frequency domain to enhance visual representations [21]. Frequency-domain analysis has also been widely explored to mitigate aliasing effects caused by downsampling. Blur pooling restores shift invariance [22], while adaptive low-pass filtering improves model generalization [23]. In the Fourier domain, FrequencyLowCut pooling removes high-frequency components beyond the Nyquist limit to reduce overfitting [24]. Recent studies further link aliasing artifacts to boundary errors and propose dealiasing filters and frequency-mixing strategies to suppress artifact-related spectral components [20]. Beyond Fourier-based methods, wavelet-based approaches provide multi-band representations for multi-scale analysis. Wavelet-integrated CNNs improve robustness to noise [25], wavelet pooling offers an alternative to standard spatial pooling [26], and Haar wavelet-based downsampling preserves fine-grained details in dense prediction tasks [27].

Prior studies primarily enhance image representations using fixed or learnable low- and high-pass filters. In contrast, our approach models long-range dependencies by learning a mask that emphasizes discriminative spectral responses. Although Global Filter Networks [14] introduce learnable Fourier-domain filters to capture long-range dependencies and replace attention layers in vision transformers to reduce computational complexity, we propose a lightweight plug-and-play module for U-Net that implicitly and efficiently captures both global context and local details, thereby improving segmentation accuracy.

3 Proposed Method

3.1 U-Net Structure with AFA Module

The U-Net architecture consists of two symmetric parts: an encoder and a decoder, connected via skip connections. The encoder progressively extracts high-level semantic and contextual features while reducing spatial resolution. At each stage $i \in \{1, 2, \ldots, N\}$:

$$\boldsymbol{E}_i = e_i(\text{Down}(\boldsymbol{E}_{i-1})), \quad \boldsymbol{E}_0 = \boldsymbol{X}, \tag{1}$$

where $e_i(\cdot)$ denotes convolutional blocks, $\text{Down}(\cdot)$ is a downsampling operator, $\boldsymbol{X}$ is the medical image input and $\boldsymbol{E}_i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times C_i}$ is the feature map at encoder stage i.

The decoder progressively upsamples the latent representation to restore spatial resolution and recover structural details. At each stage $i \in \{N-1, N-2, \ldots, 0\}$:

$$\boldsymbol{D}_i = d_i(\text{Up}([\boldsymbol{D}_{i+1}, \boldsymbol{E}_i])), \tag{2}$$

where $\boldsymbol{D}_{i+1}$ is the higher-level feature from the previous decoder stage, $\text{Up}(\cdot)$ denotes an upsampling operation, $[\cdot, \cdot]$ represents channel-wise concatenation with the corresponding encoder feature $\boldsymbol{E}_i$ (the skip connection), $d_i(\cdot)$ is the convolutional refinement block at decoder level i.

As illustrated in Fig. 1, the Adaptive Frequency Attention (AFA) module is integrated into the encoder along with a residual connection to refine the contextual representations of feature maps $\boldsymbol{E}_i$, thereby facilitating more effective information flow across subsequent encoder and decoder stages. In the

U-Net architecture augmented with AFA, for each encoder stage $i \in \{1, 2, \ldots, N\}$, the operation is defined as:

$$\hat{\boldsymbol{E}}_i = \lambda \bar{\boldsymbol{E}}_i + (1 - \lambda)\tilde{\boldsymbol{E}}_i, \quad \hat{\boldsymbol{E}}_0 = \boldsymbol{E}_0, \tag{3}$$

where $\tilde{\boldsymbol{E}}_i = e_i(\text{Down}(\hat{\boldsymbol{E}}_{i-1}))$ represents the standard convolutional feature, λ is the residual weight of AFA module and $\bar{\boldsymbol{E}}_i = \text{AFA}(\tilde{\boldsymbol{E}}_i)$ is the output produced by the AFA module.

For the decoder, at each stage $i \in \{N-1, N-2, \ldots, 0\}$:

$$\hat{\boldsymbol{D}}_i = d_i(\text{Up}([\hat{\boldsymbol{D}}_{i+1}, \hat{\boldsymbol{E}}_i])). \tag{4}$$

Therefore, the final segmentation output is obtained by:

$$\hat{\boldsymbol{Y}} = \sigma(\boldsymbol{W}_0 * \hat{\boldsymbol{D}}_0 + \boldsymbol{b}_0). \tag{5}$$

where $\boldsymbol{W}_0$ and $\boldsymbol{b}_0$ are the output convolution weights and bias, and $\sigma(\cdot)$ is activation function.

3.2 AFA Module

In AFA, we introduce a ring-shaped Gaussian mask parameterized by learnable center frequency and bandwidth. Given a feature map of spatial size $H_i \times W_i$, we first define the normalized radial frequency as

$$r(u, v) = \sqrt{\left(\frac{u - H_i/2}{H_i}\right)^2 + \left(\frac{v - W_i/2}{W_i}\right)^2}, \tag{6}$$

where (u, v) denotes the frequency coordinates after shifting the zero-frequency component to the spectrum center. The ring-shaped Gaussian mask is then defined as:

$$M_c(u, v) = \exp\left(-\frac{(r(u, v) - \mu_c)^2}{2\sigma_c^2}\right), \tag{7}$$

where c represents the channel, μ_c is the learnable center frequency, and σ_c is the learnable bandwidth.

Given the frequency feature map $\tilde{\boldsymbol{\mathcal{F}}}_i(u, v, c)$ obtained by applying DFT to the spatial feature $\tilde{\boldsymbol{E}}_i(h, w, c)$ at the ith encoder stage, where (u, v) denote frequency coordinates and (h, w) denote spatial coordinates. The spectrum extracted using the learned mask is defined as:

$$\bar{\boldsymbol{\mathcal{F}}}_i(u, v, c) = \tilde{\boldsymbol{\mathcal{F}}}_i(u, v, c) \odot M_c(u, v), \tag{8}$$

where $\odot$ denotes element-wise multiplication, and $M_c(u, v)$ is the frequency mask for channel c.

To promote diversity in feature representations across the frequency domain, we divide channels into G groups and assign each group an independent ring-shaped Gaussian mask. Each channel c is mapped to a group index $g(c) \in \{1, \ldots, G\}$ such that channels within the same group share identical frequency attention patterns, while channels from different groups specialize in distinct spectral responses. Consequently, Eqn. (8) is reformulated as:

$$\bar{\boldsymbol{\mathcal{F}}}_i(u, v, c) = \tilde{\boldsymbol{\mathcal{F}}}_i(u, v, c) \odot M_{g(c)}(u, v), \tag{9}$$

where $M_{g(c)}(u, v)$ represents the shared frequency mask for group $g(c)$.

The output of the AFA module, $\bar{\boldsymbol{E}}_i(h, w, c)$, is then derived by applying the inverse Fourier transform to the extracted spectrum $\bar{\boldsymbol{\mathcal{F}}}_i(u, v, c)$.

Spectral Orthogonality Regularization. To encourage diversity among frequency groups, we enforce orthogonality directly at the frequency response level. We flatten each 2D response into a vector $\boldsymbol{s}_g = \text{vec}(M_{g(c)}(u, v)) \in \mathbb{R}^{|\Omega|}$, where Ω denotes the set of all frequency coordinates. The orthogonality loss is then formulated as the sum of squared cosine similarities between all group pairs:

$$\mathcal{L}_{\text{Ortho}} = \sum_{i=1}^{G-1} \sum_{j=i+1}^{G} \left(\frac{\langle \boldsymbol{s}_i, \boldsymbol{s}_j \rangle}{\|\boldsymbol{s}_i\|_2 \cdot \|\boldsymbol{s}_j\|_2}\right)^2, \tag{10}$$

where $\langle \boldsymbol{s}_i, \boldsymbol{s}_j \rangle = \sum_{(u,v)\in\Omega} M_i(u,v)M_j(u,v)$. By minimizing Eqn. (10), we ensure that different groups attend to complementary spectral regions, maximizing diversity in learned frequency representations.

Spectral Preservation Constraint. Unconstrained optimization of frequency masks can result in excessive suppression of informative spectral components, leading to information loss and degraded segmentation performance. To mitigate this, we introduce a spectral energy preservation constraint. The average suppression rate for the learned mask $M_g(u,v)$ is defined as:

$$\mathcal{S}(M_g) = \frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} \big(1 - M_g(u,v)\big), \tag{11}$$

where Ω denotes the set of all frequency indices, and $M_g(u,v) \in [0,1]$ represents the frequency retention ratio at position (u,v) for group g.

To prevent excessive spectral suppression during training, we enforce a hard constraint by dynamically rescaling the suppression component when $\mathcal{S}(M_g)$ exceeds a predefined threshold $\rho \in (0,1)$:

$$M'_g(u,v) = \begin{cases} M_g(u,v), & \text{if } \mathcal{S}(M_g) \le \rho \\ 1 - \frac{\rho}{\mathcal{S}(M_g)} \cdot (1 - M_g(u,v)). & \text{otherwise} \end{cases} \tag{12}$$

This constraint is applied during forward propagation before computing the masked spectrum in Eqn. (9), ensuring that $\mathcal{S}(M_g) \le \rho$ for all groups $g \in \{1,\dots,G\}$. By maintaining at least $(1-\rho)$ fraction of spectral energy throughout training, this mechanism prevents excessive information loss while preserving gradient flow for effective optimization.

3.3 Overall Loss Function

We employ a segmentation loss $\mathcal{L}_{\text{Seg}}$, which combines the Binary Cross-Entropy (BCE) loss and the Dice loss as in [2], applied to the network output $\hat{\boldsymbol{Y}}$ in Eqn. (5). The overall loss function is formulated as:

$$\mathcal{L}_{\text{Overall}} = \mathcal{L}_{\text{Seg}} + \mathcal{L}_{\text{Ortho}}. \tag{13}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our method on five challenging medical image segmentation benchmarks: ISIC-2017 [28] (dermoscopic skin lesions), Kvasir-Seg [29] (gastrointestinal polyps), DRIVE [30] (retinal vessels), DDTI [31] (thyroid nodules in ultrasound), and BUSI [32] (breast lesions in ultrasound). This diverse selection enables a comprehensive evaluation across imaging modalities, including photography, fundoscopy, and ultrasonography. The detail description of datasets will be provided in supplementary material. For evaluation metrics, we adopt two commonly used segmentation metrics: mean Intersection over Union (mIoU) and Dice coefficient (Dice).

4.2 Implementation Details

We adopt SwinUNet [13], nnU-Net [2] and nnWNet [6] as baseline models and integrate the proposed AFA modules into their architectures as they represent pure Transformer, pure CNN and hybrid architecture. All models are trained for 200 epochs using their official default settings, following a unified training protocol to ensure fair comparison. Input images are uniformly resized to 512×512 for consistent spatial resolution. The same data augmentation, learning rate (0.01), random seeds (0) and optimizer (Adam) are applied across all experiments. Reported results are obtained directly from model predictions without post-processing.

4.3 Main Results

We adopt U-Net [1], UNet++ [5], DeepLabv3+ [33], TransUNet [7], and UWT-Net [34] for baseline comparison. Table 1 shows that AFA can be seamlessly integrated into various U-Net variants, consistently improving their performance and achieving state-of-the-art results. These results demonstrate

Table 1: Comparison of AFA integrated into representative segmentation backbones against other baseline methods. All results are reported as percentages. Best results per group are in **bold**.

Method	Params	FLOPs	BUSI		Kvasir-Seg		ISIC-2017		DDTI		DRIVE	
			Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
U-Net [1][MICCAI'15]	31.05 M	192.64 G	74.37	62.29	71.26	60.79	78.22	68.38	58.39	44.64	73.52	58.30
UNet++ [5][TMI'20]	9.16 M	34.68 G	66.69	59.41	87.67	80.32	83.75	74.99	65.72	51.70	79.80	66.42
DeepLabV3+ [33][ECCV'18]	208.70 M	166.84 G	72.88	65.02	87.83	81.20	83.31	74.41	74.35	61.38	74.17	59.12
TransU-Net [7][MedIA'24]	105.57 M	145.23 G	72.13	64.50	91.28	85.24	72.13	60.08	69.63	57.05	78.40	64.50
UWT-Net [34][MICCAI'25]	58.27 M	91.23 G	78.20	69.14	86.76	77.29	85.01	76.91	81.07	68.49	68.56	52.16
SwinU-Net [13][ECCVW'22]	27.17 M	47.32 G	82.20	69.78	88.62	79.57	83.52	71.70	73.31	**59.86**	71.48	**61.92**
+ AFA (ours)	+0.518 K	+1.093 G	**82.85**	**70.72**	**89.44**	**80.90**	**84.59**	**73.29**	**74.67**	59.57	**73.86**	58.55
nnU-Net [2][Nat. Meth.'21]	92.47 M	120.42 G	77.50	69.12	91.36	85.25	83.77	75.44	77.56	67.72	81.05	68.17
+ AFA (ours)	+3.024 K	+2.765 G	**81.15**	**73.51**	**93.38**	**88.27**	**85.54**	**77.50**	**83.93**	**74.72**	**82.35**	**70.03**
nnWNet [6][CVPR'25]	7.05 M	15.72 G	78.26	70.61	92.01	87.72	82.52	74.11	76.87	67.26	80.82	67.84
+ AFA (ours)	+2.402 K	+3.003 G	**80.60**	**73.36**	**93.56**	**88.52**	**83.73**	**75.59**	**80.39**	**71.28**	**82.71**	**70.54**

Params: parameter count (M = millions, K = thousands); + denotes increment over base model. FLOPs: floating point operations in billions (G).

Table 2: Comparison of frequency domain methods on nnU-Net backbone. All results are reported as percentages. Best results are in **bold**.

Method	Params	FLOPs	BUSI		Kvasir-Seg		ISIC-2017		DDTI		DRIVE	
			Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
nnU-Net [2][Nat. Meth.'21]	92.47 M	120.42 G	77.50	69.12	91.36	85.25	83.77	75.44	77.56	67.72	81.05	68.17
Classic spatial attention methods												
+SE [35][CVPR'18]	+0.109 M	+0.033 G	78.10	69.59	91.99	86.05	84.22	75.89	78.38	68.37	81.46	68.78
+CBAM [36][ECCV'18]	+0.109 M	+0.102 G	78.03	69.65	92.00	85.56	84.11	75.96	78.01	68.37	81.51	68.56
+ECA [37][CVPR'20]	+0.031 K	+0.033 G	78.64	69.75	92.08	85.52	84.31	76.17	77.91	68.53	81.57	68.78
Plug-and-play frequency methods												
+Blur [22][ICML'19]	+0	+0.296 G	78.38	70.59	91.10	84.86	84.44	75.94	79.29	69.01	80.72	67.71
+FLC [24][ECCV'22]	+0	+1.381 G	78.51	70.68	91.90	86.09	84.35	76.02	67.87	56.39	81.94	69.43
+DAF [20][ICLR'24]	+0	+2.698 G	78.95	70.91	91.06	84.83	84.51	76.08	80.48	70.96	81.09	68.22
+FreqMix [20][ICLR'24]	+0.877 M	+4.216 G	79.06	70.91	91.62	85.86	83.82	75.32	77.73	67.65	80.55	67.47
+DAF&FreqMix [20][ICLR'24]	+0.877 M	+6.914 G	78.35	70.88	92.73	87.06	84.52	76.11	81.92	72.12	81.74	69.15
+AFA (ours)	+3.024 K	+2.765 G	**81.15**	**73.51**	**93.38**	**88.27**	**85.54**	**77.50**	**83.93**	**74.72**	**82.35**	**70.03**

Params: parameters in millions (M); + denotes increment over nnU-Net baseline. FLOPs: floating point operations in billions (G). +0 indicates no additional trainable parameters.

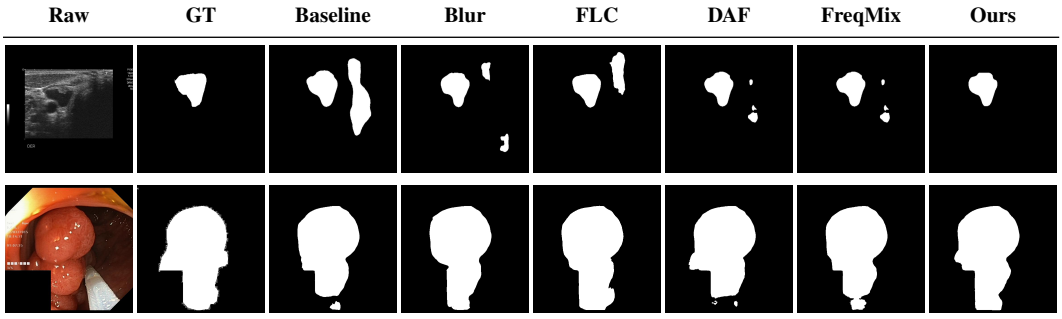


Figure 2: Qualitative comparison across different plug-and-play frequency domain methods. We use DDTI and Kvasir-Seg as visualization datasets and nnU-Net as baseline model.

that AFA effectively captures both long-range dependencies and local details, leading to improved segmentation performance. Compared with SwinU-Net, nnU-Net, and nnWNet, AFA introduces only negligible additional parameters and FLOPs.

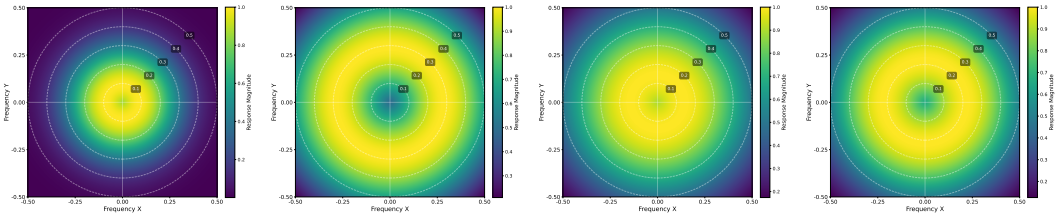
We compare AFA with other frequency-related modules, including Blur [22], FLC [24], DAF [20], and FreqMix [20]. Also, we compare AFA with classic spatial attention methods, including SE [35], CBAM [36], and ECA [37]. For the comparison, we conduct experiments on nnU-Net backbone, with additional comparisons on nnWNet provided in the supplementary material. Table 2 demonstrates that AFA consistently outperforms existing methods across all five datasets.

Table 3: Segmentation performance on DDTI with different mask ring shapes and mask weighting mechanism. “Circular” denotes the circular ring design, while “Square” refers to its square counterpart. Best results are in **bold**.

	Gaussian Mask		0/1 Mask	
	Dice↑	mIoU↑	Dice↑	mIoU↑
Circular	**83.93**	**74.72**	82.62	73.22
Square	81.97	72.38	80.63	71.31

Table 4: Segmentation performance on DDTI with different channel groupings. ”Shared” indicates that all channels share the same AFA module, whereas ”Channelwise” assigns an independent AFA to each channel. Best results are in **bold**.

AFA Grouping	Dice ↑	mIoU ↑	Params (K) ↓
Channelwise	83.01	73.52	24.192
4 channels/group	82.76	73.42	6.048
8 channels/group	**83.93**	**74.72**	3.024
16 channels/group	82.47	73.35	1.512
Shared (all channels)	82.18	72.78	**0.084**



(a) Channel group 1 (b) Channel group 2 (c) Channel group 3 (d) Channel group 4

Figure 3: Visualization of the AFA module across randomly selected channel groups on Kvasir-SEG. The results show that different channel groups in AFA learn diverse and complementary spectral patterns across feature channels.

We further visualize the segmentation output of using different plug-and-play modules. Fig. 2 presents a segmentation comparison across different plug-and-play frequency domain method using nnU-Net as the baseline. The results demonstrate that AFA reduces spurious lesion and achieves more precise localization and delivers more accurate segmentation performance compared to other methods.

4.4 Ablation Study

We conduct comprehensive ablation studies on nnU-Net to determine the optimal AFA configuration. nnU-Net serves as our baseline due to its robust performance and standardized architecture for medical image segmentation.

Mask Shape and Weighting Mechanism. Traditional frequency masks typically use square shapes with binary weighting, which introduce directional bias and discard useful spectral components. In contrast, our circular ring masks with Gaussian weighting provide isotropic frequency treatment and smoothly emphasize target bands while preserving adjacent components. As shown in Table 3, Gaussian weighting outperforms binary masks, and the circular design surpasses the square counterpart, validating their effectiveness for medical image segmentation.

Effect of Channel Grouping. We visualize the AFA module of different channel grouping and investigate the effect of channel grouping on medical image segmentation performance. As shown in Table 4, introducing channel grouping improves segmentation accuracy. However, excessively large groups lead to performance degradation. The best results are achieved when eight channels are assigned to each group, which balances both parameter costs and segmentation performances. Furthermore, Fig. 3 shows that AFA of different channel groups capture different spectral patterns across different channel groups. Additional visualization of AFA across different datasets will be provided in the supplementary material.

AFA Insertion Analysis. We systematically study the effect of placing AFA at different encoder stages. Table 5 shows that inserting AFA from stage 1 already improves performance, with accuracy increasing as more stages are equipped. The best results occur when all encoder stages use AFA, highlighting that enhancing frequency-domain representations across multiple semantic levels effectively boosts contextual understanding. In contrast, applying AFA only to the decoder performs similarly to the baseline, suggesting that AFA primarily benefits semantic encoding in the encoder.

Table 5: Performance comparison of integrating AFA into different encoder stages on DDTI. Best results are in **bold**.

AFA Insertion Strategy	Stages							Params	Performance	
	1	2	3	4	5	6	7	Δ	Dice ↑	mIoU ↑
Baseline (no AFA)	✗	✗	✗	✗	✗	✗	✗	0	77.56	67.72
Stage [1] only	✓	✗	✗	✗	✗	✗	✗	48	81.07	71.34
Stages [1,2]	✓	✓	✗	✗	✗	✗	✗	144	81.73	72.48
Stages [1,2,3]	✓	✓	✓	✗	✗	✗	✗	336	82.29	72.50
Stages [1,2,3,4]	✓	✓	✓	✓	✗	✗	✗	720	82.51	73.42
Stages [1,2,3,4,5]	✓	✓	✓	✓	✓	✗	✗	1,488	82.67	73.16
Stage [7] only	✗	✗	✗	✗	✗	✗	✓	768	81.52	72.00
Stages [6,7]	✗	✗	✗	✗	✗	✓	✓	1,536	80.62	70.78
Stages [5,6,7]	✗	✗	✗	✗	✓	✓	✓	2,304	82.78	73.46
All Encoder Stages [1–7]	✓	✓	✓	✓	✓	✓	✓	3,024	**83.93**	**74.72**
All Decoder Stages [1–6]	✓	✓	✓	✓	✓	✓	–	2,256	77.91	68.03

Table 6: Performance comparison of individual AFA components on DDTI. Best results are in **bold**.

Method	Dice ↑	mIoU ↑
nnU-Net baseline	77.56	67.72
+ $\mathcal{L}_{\text{Seg}}$	82.03	72.84
+ $\mathcal{L}_{\text{Seg}}$+ $\mathcal{S}(\mathrm{M}_g)$	82.35	73.02
+ $\mathcal{L}_{\text{Seg}}$+ $\mathcal{L}_{\text{Ortho}}$	82.64	73.25
+ $\mathcal{L}_{\text{Seg}}$ + $\mathcal{L}_{\text{Ortho}}$ + $\mathcal{S}(\mathrm{M}_g)$	**83.93**	**74.72**

Table 7: Ablation study on suppression rate ρ and residual weight λ in AFA module.

	Suppression Rate (ρ)					Residual Weight (λ)				
Value	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
DICE (%)	80.96	**83.93**	81.81	81.80	80.89	80.95	**83.93**	83.36	78.37	78.89

Effect of AFA Components. We evaluate the individual contributions of each AFA component on the DDTI dataset. As shown in Table 6, incorporating the learnable mask enhances segmentation performance. Further adding the orthogonality regularization loss and spectral preservation constraint enables AFA to learn more diverse masks, leading to additional improvements in medical image segmentation accuracy.

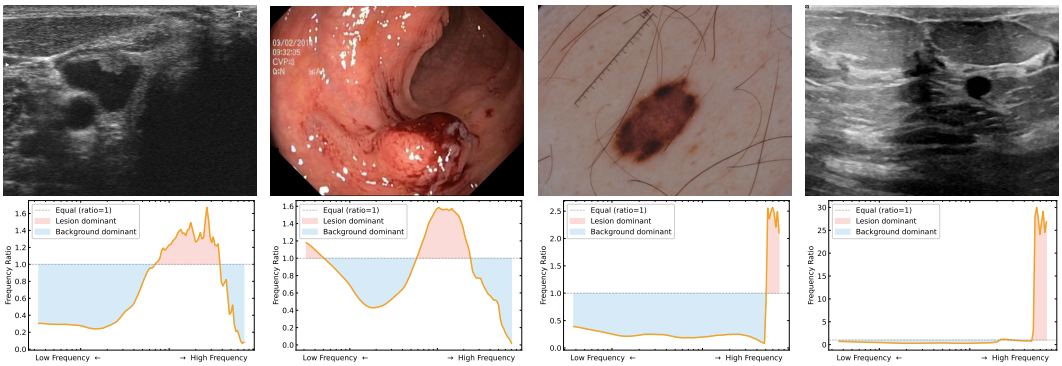
Effect of Hyperparameters. We evaluate the impact of the residual weight λ and suppression threshold ρ. As shown in Table 7, performance improves with increasing ρ and λ up to a point, but declines when values are too large, as either frequency suppression becomes insufficient or the model over-relies on filtered features. Optimal performance is achieved with $\lambda = \rho = 0.3$.

4.5 Semantic Context Study

As shown in Table 1, the proposed AFA module consistently improves not only nnU-Net [2], a pure CNN architecture, but also SwinUNet [13], a pure Transformer model, as well as nnWNet [6], a hybrid architecture. We hypothesize that AFA not only effectively captures both long-range dependencies and local details, but also enhances semantic representation through frequency-domain modeling given that U-Net-based architectures rely on semantic context to distinguish regions.

To further investigate this, we conduct a systematic frequency analysis using the Fourier Transform, which maps images from the spatial domain to the frequency domain. To quantify spectral differences between lesional and normal regions, we define a frequency dominance ratio $R(f)$ (details can be found in the appendix), where $R(f) > 1$ indicates dominance of lesional regions at frequency f, and $R(f) < 1$ indicates dominance of normal regions. As shown in Fig. 4, lesional regions exhibit distinct spectral characteristics compared to normal regions. These findings suggest that AFA can capture semantic context in the frequency domain, thereby enhancing feature representations and improving segmentation performance.

We conduct interpretability studies on BUSI, ISIC-2017, DRIVE, and Kvasir-Seg to examine whether AFA guides the model to focus on lesions. Using Grad-CAM [38] with nnU-Net tools [39], we



(a) DDTI dataset (b) Kvasir-Seg dataset (c) ISIC 2017 dataset (d) BUSI dataset

Figure 4: Frequency analysis of medical images. When the frequency dominance ratio $R(f) > 1$, lesional regions dominate, whereas $R(f) < 1$ indicates dominance of normal regions. This demonstrates that lesional regions exhibit distinct spectral patterns from normal regions across specific frequency bands.

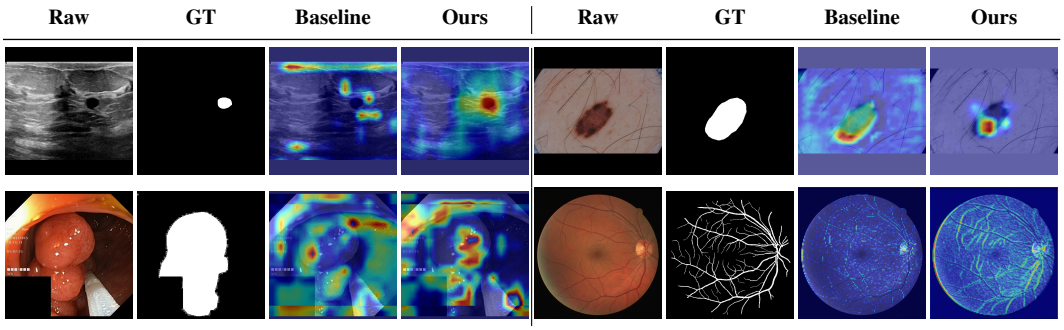


Figure 5: Grad-CAM visualization results for interpretability study of AFA module. We compare the visualization results of AFA-integrated network (nnU-Net+AFA) with baseline (nnU-Net).

visualize the last convolutional layer for BUSI, ISIC-2017, and Kvasir-Seg, and the third intermediate layer for DRIVE, as fine-grained vessel structures are better captured there. Layer-wise analysis confirms the third layer as most informative (final layer results in supplementary material). Compared to baselines, AFA substantially improves lesion localization and alignment with ground truth. Fig. 5 shows that baseline models exhibit misactivation or incorrect activation regions, while the AFA module improves lesion localization and consistency with ground truth.

5 Conclusion

In this work, we propose Adaptive Frequency Attention (AFA), a plug-and-play module that enhances U-Net architectures by simultaneously modeling long-range dependencies and local details in the frequency domain. AFA employs a learnable ring-shaped Gaussian mask to selectively emphasize discriminative spectral components, thereby improving contextual representations. To promote diverse frequency responses, feature channels are partitioned into groups with an orthogonality-based regularization. In addition, a spectral preservation constraint regulates suppression strength to prevent excessive information loss. Extensive experiments demonstrate that AFA consistently improves U-Net variants, achieving state-of-the-art performance across multiple medical image segmentation benchmarks. Furthermore, frequency-domain analysis suggests that AFA captures semantic-related spectral patterns, contributing to improved segmentation performance.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical*

Image Computing and Computer-Assisted Intervention (MICCAI), pages 234–241, 2015.

[2] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[3] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging (TMI)*, 37(12):2663–2674, 2018.

[4] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2018.

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

[6] Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 20852–20862, 2025.

[7] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, page 103280, 2024.

[8] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059, 2020.

[9] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] 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. In *International Conference on Learning Representations*, 2021.

[12] 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. *International Conference on Learning Representations (ICLR)*, 2021.

[13] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 205–218, 2022.

[14] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 980–993, 2021.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[16] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 488–498, 2024.

[17] Antonio A Abello, Roberto Hirata, and Zhangyang Wang. Dissecting the high-frequency bias in convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 863–871, 2021.

[18] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:1256–1272, 2021.

[19] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

[20] Linwei Chen, Lin Gu, and Ying Fu. When semantic segmentation meets frequency aliasing. In *International Conference on Learning Representations (ICLR)*, 2024.

[21] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 783–792, 2021.

[22] Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, pages 7324–7334, 2019.

[23] Xueyan Zou, Fanyi Xiao, Zhiding Yu, Yuheng Li, and Yong Jae Lee. Delving deeper into anti-aliasing in convnets. *International Journal of Computer Vision (IJCV)*, 131(1):67–81, 2023.

[24] Julia Grabinski, Steffen Jung, Janis Keuper, and Margret Keuper. Frequencylowcut pooling-plug and play against catastrophic overfitting. In *European Conference on Computer Vision (ECCV)*, pages 36–57, 2022.

[25] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7245–7254, 2020.

[26] Travis Williams and Robert Li. Wavelet pooling for convolutional neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[27] Guoping Xu, Wentao Liao, Xuan Zhang, Chang Li, Xinwei He, and Xinglong Wu. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern recognition (PR)*, 143:109819, 2023.

[28] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 168–172, 2018.

[29] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *Proceedings of the International Conference on MultiMedia Modeling (MMM)*, pages 451–462, 2020.

[30] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging (TMI)*, 23(4):501–509, 2004.

[31] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *International symposium on medical information processing and analysis*, volume 9287, pages 188–193. SPIE, 2015.

[32] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

[33] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[34] Pengcheng Zhang, Xiaocao Ouyang, and Ran Peng. Uwt-net: Mining low-frequency feature information for medical image segmentation. pages 615–624, 2025.

[35] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[37] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[39] Youssef Abuzeid, Shimaa El-Bana, and Ahmad Al-Kabbany. Xai-driven diagnosis of generalization failure in state-space cerebrovascular segmentation models: A case study on domain shift between rsna and topcow datasets, 2025.

A Technical appendices and supplementary material

A.1 Lesional Region Frequency Analysis

To explore the intrinsic spectral differences between lesional and normal regions, we perform a systematic frequency analysis based on Fourier Transform, which converts images from the spatial domain to the frequency domain. Given an image $\boldsymbol{I}(x, y)$ of size $H \times W$, the two-dimensional Discrete Fourier Transform (DFT) is defined as:

$$\boldsymbol{F}(u, v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \boldsymbol{I}(x, y) e^{-j2\pi\left(\frac{ux}{H} + \frac{vy}{W}\right)}, \tag{14}$$

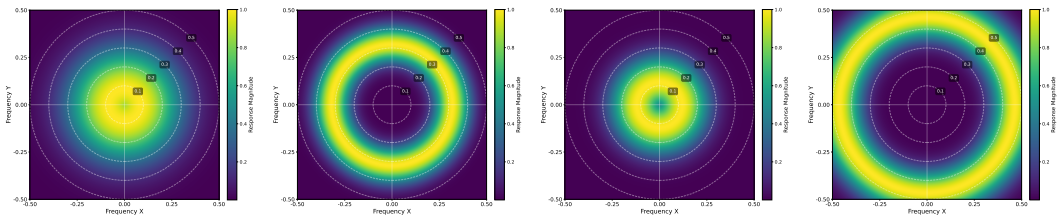
where (x, y) denotes the spatial coordinates, (u, v) denotes the frequency coordinates, $j = \sqrt{-1}$, and $\boldsymbol{F}(u, v)$ denotes the complex Fourier coefficient that encodes amplitude and phase.

Then, the inverse transform reconstructs the image:

$$\boldsymbol{I}(x, y) = \frac{1}{H \times W} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \boldsymbol{F}(u, v) e^{j2\pi\left(\frac{ux}{H} + \frac{vy}{W}\right)}. \tag{15}$$

Our analysis is grounded in pixel-level ground truth annotations and employs rigorous spectral processing techniques to extract reliable frequency signatures. The analysis pipeline comprises three stages: patch extraction, frequency domain analysis and statistical aggregation.

Patch Extraction. Using binary ground truth mask, we extract patches from both lesional and normal regions. To avoid contamination, a minimum margin of 2 pixels is maintained from the boundaries of lesional areas. All patches undergo standardized preprocessing: RGB images are converted to luminance, which is then normalized to zero mean and unit variance. To reduce spectral leakage artifacts, a Hann window is applied to each normalized luminance patch $\bar{\boldsymbol{P}}(x, y)$:



(a) AFA on Kvasir-Seg (b) AFA on DDTI (c) AFA on DRIVE (d) AFA on ISIC-2017

Figure 6: Visualization of AFA module on frequency domain. Higher brightness corresponds to stronger frequency emphasis. The results indicate that AFA can attend to different frequency bands depending on the dataset.

$$\hat{\boldsymbol{P}}(x,y) = \bar{\boldsymbol{P}}(x,y) \cdot \sin^2\left(\frac{\pi x}{H_p}\right) \cdot \sin^2\left(\frac{\pi y}{W_p}\right), \tag{16}$$

where H_p and W_p denote the height and width of the patch, respectively.

Frequency Domain Analysis. The Energy Spectral Density (ESD) is derived by applying the DFT in Eqn. (14):

$$\text{ESD}(u,v) = \frac{|\hat{\boldsymbol{F}}(u,v)|^2}{H_p \cdot W_p}, \tag{17}$$

where $\hat{\boldsymbol{F}}(u,v)$ represents the complex Fourier coefficient of $\hat{\boldsymbol{P}}(x,y)$, and $|\cdot|^2$ denotes the squared magnitude. Normalization by the total number of pixels ensures that ESD values are comparable across patches of different sizes.

We convert the 2D frequency coordinates (u,v) into discrete radial frequency indices:

$$r(u,v) = \sqrt{\left(\frac{u - H_p/2}{H_p}\right)^2 + \left(\frac{v - W_p/2}{W_p}\right)^2}. \tag{18}$$

Then, we aggregate the ESD along the discrete radial axis. The radial ESD corresponding to a discrete frequency index f is defined as:

$$\mathcal{E}(f) = \frac{1}{|\mathcal{B}(f)|} \sum_{(u,v)\in\mathcal{B}(f)} \text{ESD}(u,v), \tag{19}$$

where $\mathcal{B}(f)$ denotes the set of frequency coordinates (u,v) whose discrete radial frequency equals f:

$$\mathcal{B}(f) = \{(u,v) \mid r(u,v) = f\}. \tag{20}$$

This discrete radial aggregation yields a rotation-invariant spectral signature for comparing lesional and normal regions.

Statistical Aggregation. For lesional regions, the dataset-level radial ESD $\bar{\mathcal{E}}_p(f)$ is obtained by averaging $\mathcal{E}_p(f)$ over all lesional patches. Similarly, $\bar{\mathcal{E}}_n(f)$ for normal regions is computed in the same way. To quantify the relative spectral prominence of lesional versus normal regions, we define the frequency dominance ratio $R(f)$ as:

$$R(f) = \frac{\bar{\mathcal{E}}_p(f)}{\bar{\mathcal{E}}_n(f)}, \tag{21}$$

where $R(f) > 1$ indicates lesional region dominance at frequency f while $R(f) < 1$ indicates normal region dominance at frequency f. As shown in Fig. 4, lesional regions exhibit distinct spectral characteristics compared to normal regions.

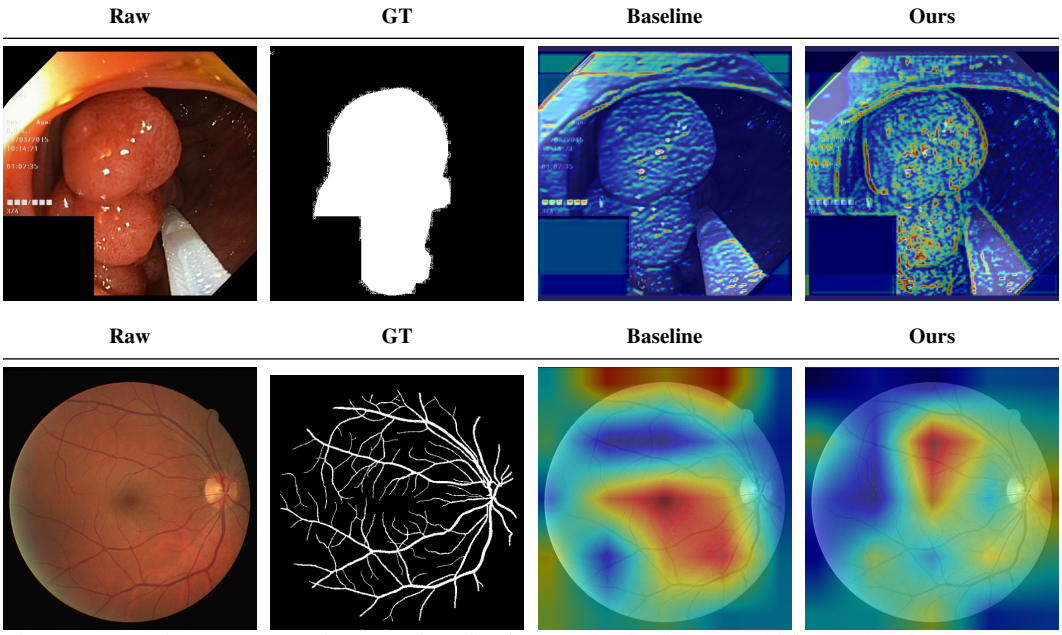


Figure 7: Supplementary Grad-CAM visualization results for interpretability study of AFA module. We compare the visualization results of AFA-integrated network (nnU-Net+AFA) with baseline (nnU-Net). Row 1 represents the third intermediate layer of the model trained on Kvasir-Seg dataset, while row 2 represents the last convolutional layer of the model trained on DRIVE dataset.

A.2 Additional Visualizations

We visualize the AFA module across different datasets. As shown in Fig. 6, while different types of lesional regions show different frequency patterns, the visualization results reveal that the AFA module effectively suppresses signals from normal regions while amplifying frequency-domain information associated with lesional regions. These visualization results further validate that AFA could capture the lesion-dominated frequency band.

We extend the Grad-CAM visualization to include the third intermediate layer of the model trained on the Kvasir-Seg dataset and the last convolutional layer of the model trained on the DRIVE dataset. As illustrated in Fig 7, the visualizations reveal that AFA not only enriches semantic feature representations but also sharpens the model's spatial focus toward fine-grained lesion structures. Notably, the enhanced intermediate-layer activations demonstrate that AFA facilitates more precise localization of lesion boundaries and subtle pathological details, underscoring its effectiveness in guiding the network to attend to clinically relevant regions across both shallow and deep feature hierarchies.

A.3 Additional Experimental Results

To further validate the generalizability of AFA, we provide additional comparisons on the nnWNet backbone across all five 2D benchmarks. As shown in Table 8, AFA consistently achieves the best performance across all datasets when integrated with nnWNet, outperforming competing frequency-domain methods including Blur, FLC, DAF, and FreqMix. This is consistent with the results reported on nnU-Net in the main paper, demonstrating that AFA is architecture-agnostic and generalizes well across both pure CNN and hybrid CNN-Transformer backbones. AFA leverages learnable Gaussian masks to selectively amplify task-relevant frequency bands, resulting in more discriminative feature representations and consistently superior segmentation performance.

A.4 Dataset Details

We evaluate our method on five diverse medical image segmentation benchmarks, each presenting unique clinical challenges and imaging characteristics.

Table 8: Comparison of different frequency-based methods on nnWNet.

Method	BUSI		Kvasir-Seg		ISIC-2017		DDTI		DRIVE	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
nnWNet [6][CVPR'25]	78.26	70.61	92.01	87.72	82.52	74.11	76.87	67.26	80.82	67.84
+Blur [22][ICML'19]	79.48	72.00	91.10	84.85	81.76	73.19	79.72	70.09	82.02	69.56
+FLC [24][ECCV'22]	78.63	71.39	92.32	86.52	83.27	74.96	78.95	70.46	81.95	69.46
+DAF [20][ICLR'24]	78.96	71.65	92.57	87.07	83.13	74.63	77.36	67.82	81.12	68.26
+FreqMix [20][ICLR'24]	78.48	70.93	92.85	87.39	83.36	74.83	77.98	68.15	81.89	69.37
+DAF&FreqMix [20][ICLR'24]	78.35	70.88	93.25	87.90	83.34	74.87	80.25	**71.42**	81.74	69.15
+AFA (ours)	**80.60**	**73.36**	**93.56**	**88.52**	**83.73**	**75.59**	**80.39**	71.28	**82.71**	**70.54**

BUSI [32] comprises breast ultrasound images annotated for benign and malignant tumor segmentation. Key challenges include characteristic speckle noise, variable lesion morphology with acoustic shadowing, and the need to differentiate tumors from acoustically similar anatomical structures.

Kvasir-Seg [29] consists of gastrointestinal endoscopy images for polyp segmentation. As polyps are precursors to colorectal cancer, accurate delineation is clinically essential. The dataset features significant diversity in polyp size, shape, and appearance under varying illumination conditions, posing substantial challenges for precise localization.

ISIC-2017 [28] contains dermoscopic images for skin lesion segmentation, a critical task in computer-aided melanoma diagnosis. The dataset exhibits challenges including irregular lesion boundaries, high intra-class variation in color and texture, and low contrast between lesions and healthy skin.

DRIVE [30] is a widely-used benchmark for retinal vessel segmentation from fundus photographs. Accurate vessel mapping is crucial for diagnosing systemic diseases such as diabetes and hypertension. The primary difficulties arise from the complex tree-like vascular topology, presence of pathological artifacts, and subtle contrast between thin vessels and the retinal background.

DDTI [31] provides thyroid ultrasound images for nodule segmentation, essential for cancer risk stratification. The dataset is challenging due to inherent speckle noise, heterogeneous echogenicity patterns, and ill-defined nodule boundaries within surrounding tissue.

A.5 Discussion

Limitations. Although the proposed method achieves strong performance across extensive 2D medical image segmentation tasks, the current AFA design processes features isotropically, which may limit its effectiveness in 3D scenarios, where medical volumes often exhibit anisotropic structures and varying slice-wise spectral characteristics. Therefore, future work will focus on extending AFA to 3D settings to better model anisotropic spectral representations and further enhance segmentation performance in 3D medical imaging.

Impacts. This work has important implications for medical image segmentation. The proposed method is designed as a lightweight plug-and-play module for U-Net-based architectures, consistently improving the segmentation performance of different U-Net variants. Due to its flexibility and compatibility, the module can be easily integrated into existing medical image segmentation systems and applications built upon U-Net-based architectures.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction accurately reflect the contributions and scope of the paper, clearly describing both the proposed methodological innovations and the experimental findings.

   Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not include formal theoretical results such as theorems or lemmas. The method is motivated by empirical observations and supported by experimental evaluation. All mathematical formulations (DFT, Gaussian mask, orthogonality loss, spectral preservation constraint) are fully defined with equations numbered and cross-referenced in Section 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The information needed to reproduce the main experimental results is introduced in Section 3, Section 4, and the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and model weights are not released at submission time to preserve anonymity. However, all five evaluation datasets (BUSI, Kvasir-Seg, ISIC-2017, DRIVE, DDTI) are publicly available, and all baseline implementations follow their official repositories.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and test details can be seen in Section 4 and the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experiment did not include error bars, confidence intervals, or other statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sufficient information on the computer resources can be seen in Section 4 and the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper fully compliant with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the potential societal impacts of the work in the supplementary material.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects. All experiments use existing, publicly released, de-identified medical imaging datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects. All data used are publicly available, pre-existing, anonymized medical imaging datasets (BUSI, Kvasir-Seg, ISIC-2017, DRIVE, DDTI). No new patient data was collected, and IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.