Ultrasound-Former: An Efficient Transformer based Method for Breast Lesions Segmentation in Ultrasound Images

Zhaohong Huang[1], Pengfei Jiang[2], Jiancong Zheng[2], Jiarui Wu[1], Yu Yang[1]

[1]Institute of Artificial Intelligence, Xiamen University
[2]School of Informatics, Xiamen University

Deep Learning(AI) 36920231153197, 31520231154278, 23020231154261, 36920231153240, 36920231153255

Abstract

Breast tumor segmentation of ultrasound images provides valuable information of tumors for early detection and diagnosis. However, speckle noise and blurred boundaries present challenges for breast lesions segmentation, especially for malignant tumors with irregular shapes. Recent vision transformers have shown promising performance in handling the variation through global context modeling. Still, they have not thoroughly solved the problem of ambiguous boundaries as they ignore the complementary usage of the boundary knowledge. In this paper, we propose an efficient transformer based method, Ultrasound-Former, to simultaneously address speckle noise interference and boundary problems of breast lesion segmentation. Specifically, we propose two modules: the Noise Suppression Module (NSM) and the Boundary Refinement Module (BRM). The NSM filters noise information from the coarse-grained feature maps, while the BRM progressively refines the boundaries of significant lesion objects, effectively optimizing blurred boundaries. Through extensive experiments on the breast ultrasound dataset, we demonstrate that Ultrasound-Former outperforms state-of-the-art methods for medical image analysis.

Introduction

Breast cancer is a common female disease, which seriously threatens women's health and life. Therefore, regular breast screening and diagnosis are very important to formulate treatment plans and improve survival rates. Due to the flexibility and convenience of ultrasound imaging, it has become a convention modality for breast tumors screening. In recent years, many deep learning methods based on ultrasound images have been proposed for breast lesion segmentation. However, complex ultrasound patterns continue to pose the following challenges: 1). Blurred boundaries caused by low contrast between foreground and background, 2). Segmentation disruption due to speckle noise (as illustrated in Figure 1).

To further address the issue of blurred boundaries caused by breast lesions, two optimization strategies: expanding the receptive field and the attention mechanisms have been widely used. The dilated convolution operation is a commonly used strategy to expand the receptive field. For example, (Hu et al. 2019) obtained the large receptive field of breast tumors by using dilated convolutions in deeper network layers. In terms of attention mechanism, (Lee, Park, and Hwang 2020) proposed a channel attention module to further improve the performance of U-Net for breast lesions segmentation. (Yan et al. 2022) proposed an attention enhanced U-Net with hybrid dilated convolution, merging dilated convolutions with an attention mechanism. Although progress has been made by these methods, the optimization paradigm from fine to coarse granularity struggles to capture prominent object regions in deeper convolutional layers, where object regions and boundaries stand as two crucial distinguishing features between normal tissue and breast tumors. Thus, we propose an iteratively enhanced Boundary Refinement Module (BRM) based on a global map. Our motivation stems from the fact that, during breast tumor annotation, clinicians first roughly locate a lesion area and then accurately extract its silhouette mask according to the local features. Within the Ultrasound-Former, we predict the coarse region first and subsequently model the boundaries implicitly through axial reverse attention. There are three advantages to this strategy, including better learning ability, improved generalization capability, and higher training efficiency.

In ultrasound imaging, speckle noise significantly impacts segmentation accuracy by propagating across various convolutional layers at different scales. Current methods primarily leverage the concept of deep supervision to develop refined networks (Qi, Wu, and Chan 2023), exploring neighboring decisions to correct potential errors induced by speckle

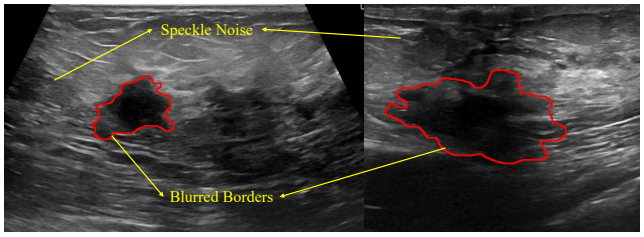

Figure 1: Challenges of breast ultrasound image segmentation task. The red line are boundaries of the breast lesions. 1) severe speckle noise are in BUS images and 2) The boundaries of breast tumors are blurred.

Table 1: The network's performance variation when eliminating high-frequency information. We use the mainstream method UNet (Ronneberger, Fischer, and Brox 2015) to evaluate the impact of high-frequency on ultrasound image segmentation on BUSI testset (Al-Dhabyani et al. 2020). Building upon (Dong, Wang, and Wang 2023), Low-Pass Filter consists of multiple pooling operations. We integrated Low-Pass Filter into the last two stages of the UNet architecture. ↑ denotes higher the better and ↓ denotes lower the better.

Method	mDice↑	mIoU↑	MAE↓
UNet	0.7023	0.6073	0.0509
UNet + Low-Pass Filter	0.7546(+5.23)	0.6579(+5.06)	0.0421(-0.88)

noise. However, we propose addressing noise influence from a more fundamental perspective by introducing "frequency." In an intriguing experiment, we examined the network's performance variation when eliminating high-frequency information (detail and noise) in deeper layers. As shown in Table 1, we observed a substantial improvement in model performance when the network included only low-pass operators (solely containing low-frequency information), indicating that speckle noise within high-frequency information disrupts spatial consistency. To address this phenomenon, we introduce a Noise Suppression Module, decoupling high and low-frequency information in feature maps and denoising the high-frequency components. While, following prior work's principles, Ultrasound-Former also incorporates a deep supervision mechanism.

Our method, built upon Transformer-based encoder, BRM and NSM modules for breast lesions segmentation in ultrasound image segmentation, dubbed Ultrasound-Former, is elucidated in Figure 2. Its efficacy is validated through extensive experiments on the breast ultrasound dataset and results hightlight significantly improvement over existing methods. Our contributions include:

- We present a novel breast lesions segmentation framework, termed Ultrasound-Former. Unlike existing CNN-based methods, we adopt the pyramid vision transformer as an encoder to extract more robust features.
- To support our framework, we introduce two simple modules. Specifically, NSM is utilized to suppress speckle noise within high-frequency information, while BRM performs boundary refinement based on coarse regions.
- Comparative experiments juxtaposed with leading-edge medical image segmentation models demonstrate the superior efficacy of our method on the breast ultrasound dataset.

Related Work

Ultrasonic image segmentation

The powerful nonlinear learning ability makes full convolution network (FCN) and U-Net have achieved great success in medical images segmentation (Zhou et al. 2018; Huang et al. 2020). Enlightened by this, many deep learning methods are proposed to segment breast lesions from ultrasound images. In 2018, (Almajalid et al. 2018) are the first to systematically evaluate the impact of different FCN variants on breast lesions segmentation and achieve segmentation results that outperform traditional methods. MDF-Net (Qi, Wu, and Chan 2023) introduces a novel multi-scale dynamic fusion strategy, employing a two-stage end-to-end architecture to achieve enhanced feature exploration and noise reduction. NU-Net (Chen et al. 2022) utilizes sub-networks of varying depths with shared weights to attain robust representations of breast tumors.

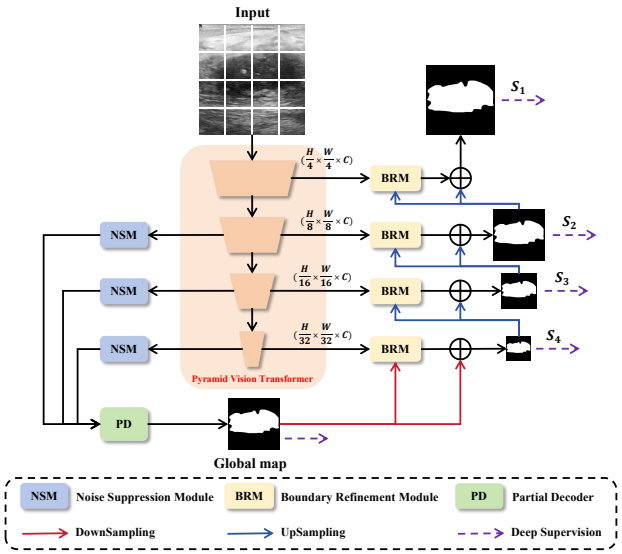

Figure 2: The framework of our proposed Ultrasound-Former comprises primarily of the Pyramid Vision Transformer, Partial Decoder (Fan et al. 2020), Noise Suppression Module, and Boundary Refinement Module.

Vision Transformers

Since 2017, the Transformer proposed by (Vaswani et al. 2017). This method based on attention mechanism and completely eliminating convolution has attracted the attention of scholars. Later, the proposal of Vision Transformer (ViT) (Dosovitskiy et al. 2020) introduced Transformer into the computer vision field for the first time, and ViT validated the feasibility of pure transformer architectures for computer vision tasks. Although the transformers are originally proposed to explore global dependency, recent studies find that the transformers also need local communication (Wang et al. 2021; Cao et al. 2022), which can be achieved through the local window shift or pyramid architecture. As for medical image segmentation, the effectiveness of vision transformers is verified by TransUNet (Chen et al. 2021) and TransFuse (Zhang, Liu, and Hu 2021).

The Proposed Method

An overview of Ultrasound-Former is shown in Figure 2. Upon inputting a ultrasound image, we initially extract four

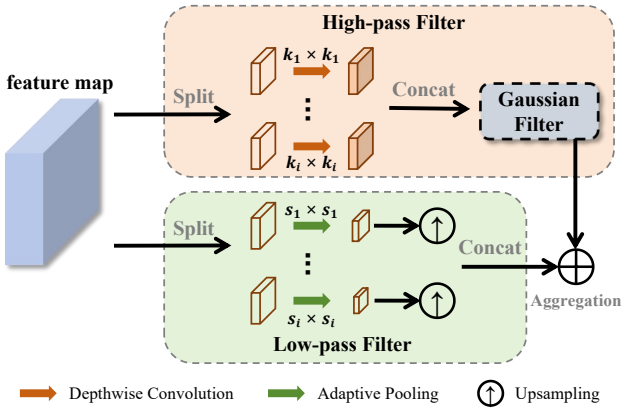

Figure 3: Overall architecture of NSM.

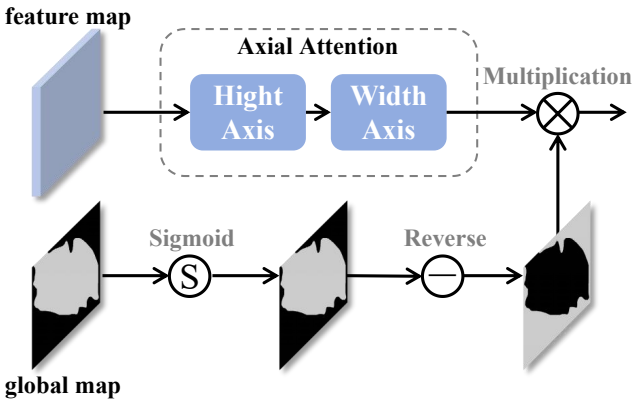

Figure 4: Overall architecture of BRM.

levels of feature maps of various scales sequentially utilizing the Pyramid Vision Transformer (PVT) block (Wang et al. 2021). We input the feature maps from the last three stages into NSM individually for the suppression of speckle noise, followed by utilizing parallel partial decoder (Fan et al. 2020) to generate high-level semantic global maps. Lastly, a set of reverse axial attention mechanisms is employed to refine the lesion boundaries progressively. Detailed expositions of NSM and BRM are presented as follows.

Noise Suppression Module

Speckle noise, a complex physical characteristic in ultrasound images, often leads to confusion in object localization. The frequency representation can be used as a new paradigm of learning difference between categories, which can excavate the information ignored by human vision. To mitigate this, we propose a Noise Suppression Module (NSM) which consider speckle noise suppression from a frequency perspective, as illustrated in Figure 3.

Low-pass Filter (LPF). Low-frequency components occupy most of the energy in the absolute image and represent most of the semantic information. A low-pass filter allows signals below the cutoff frequency to pass, while signals above the cutoff frequency are obstructed. Thus, we employ typical average pooling as a low-pass filter. However, the cutoff frequencies of different images are different. To this end, we employ channel split, the feature map is partitioned into multiple groups and control different kernels and strides in multigroups to generate low-pass filters. For m-th group, we have:

$$LPF_m\left(v^m\right) = Up\left(\Gamma_{s\times s}\left(v^m\right)\right), \tag{1}$$

where $Up\left(\cdot\right)$ represents upsampling and $\Gamma_{s\times s}$ denotes the adaptive average pooling with the output size of $s \times s$.

High-pass Filter (HPF). High-frequency information is crucial to preserve details in segmentation. As a typical high-pass operator, convolution can filter out irrelevant low-frequency redundant components to retain favorable high-frequency components. The high-frequency components determine the image quality and the cutoff frequency of the high-pass for each image is different. Similar to LPF, we partition the feature map into n groups. For each group, we use a convolution layer with different kernels to simulate the cutoff frequencies in different high-pass filters. For the n-th group, we have:

$$HPF_n\left(v^n\right)) = \Lambda_{k\times k}\left(v^n\right), \tag{2}$$

where $\Lambda_{k\times k}$ denotes the depthwise convolution layer with kernel size of $k \times k$. The continuous accumulation of speckle noise within the internal high frequencies often yields adverse effects on the extracted high-frequency information. Therefore, we employ Gaussian Filtering on the high-frequency features to effectively eliminate noise.

$$W(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{3}$$

$$G(x,y) = \sum_{i=-k}^{k}\sum_{j=-k}^{k} W(i,j)\cdot I(x+i, y+j), \tag{4}$$

where $G(x,y)$ represents the value of the Gaussian function at the spatial coordinates (x,y), σ is the standard deviation of the Gaussian function, and k stands for the window size of the Gaussian filter. The final output F_{NSM} is obtained by summing the denoised high-frequency information with the low-frequency information:

$$F_{NSM} = HPF_n\left(v^n\right) + G(LPF_n\left(v^n\right)). \tag{5}$$

Boundary Refinement Module

As discussed above, our global map S_global is derived from the deepest segment of the network, achieved via partial decoders, which can only capture a relatively rough location of the breast lesion, without structural details. To address this issue, we propose Boundary Refinement Module (BRM) to progressively mine discriminative breast tumor through an erasing foreground object manner, as illustrated in Figure 4. We propose to adaptively learn the reverse attention in three parallel high-level features. In other words, our architecture can sequentially mine complementary regions and details by erasing the existing estimated lesion regions from highlevel side-output features, where the existing estimation is up-sampled from the deeper layer. Simultaneously,

Table 2: Quantitative comparison of different methods on BUSI (Al-Dhabyani et al. 2020) to validate our model's learning ability. ↑ denotes higher the better and ↓ denotes lower the better. Red indicates the best results and blue represents the second-best results.

Method	All			Benign			Malignant		
	mDice↑	mIoU↑	MAE↓	mDice↑	mIoU↑	MAE↓	mDice↑	mIoU↑	MAE↓
UNet (Ronneberger, Fischer, and Brox 2015)	0.6943	0.6033	0.0496	0.7219	0.6362	0.0380	0.6232	0.5183	0.0798
Attention U-Net (Oktay et al. 2018)	0.6934	0.6016	0.0509	0.7247	0.6374	0.0378	0.6125	0.5092	0.0845
UNet++ (Zhou et al. 2018)	0.7023	0.6070	0.0509	0.7212	0.6301	0.0398	0.6538	0.5476	0.0796
UNet3+ (Huang et al. 2020)	0.7055	0.6139	0.0493	0.7358	0.6433	0.0388	0.6487	0.5414	0.0765
PraNet (Fan et al. 2020)	0.7698	0.6847	0.0413	0.7841	0.7037	0.0320	0.7330	0.6272	0.0654
DoubleU-Net (Jha et al. 2020)	0.7735	0.6870	0.0461	0.8016	0.7179	0.0333	0.7010	0.5885	0.0790
UACANet (Kim, Lee, and Kim 2021)	0.7473	0.6650	0.0442	0.7593	0.6773	0.0353	0.7163	0.6089	0.0672
SANet (Wei et al. 2021)	0.7708	0.6842	0.0458	0.7929	0.7074	0.0351	0.7136	0.6065	0.0732
UNext (Valanarasu and Patel 2022)	0.7171	0.6258	0.0436	0.7366	0.6509	0.0332	0.6668	0.5613	0.0702
CaraNet (Lou et al. 2022)	0.7769	0.6968	0.0383	0.7947	0.7199	0.0287	0.7289	0.6267	0.0633
DuAT (Tang et al. 2022)	0.8017	0.7163	0.0406	0.8164	0.7137	0.0314	0.7285	0.6284	0.0667
XBound-Former (Wang et al. 2023)	0.7986	0.7083	0.0419	0.8059	0.7094	0.0322	0.7283	0.6215	0.0670
PVT-CASCADE (Rahman and Marculescu 2023)	0.8118	0.7270	0.0380	0.8374	0.7582	0.0245	0.7456	0.6465	0.0619
Ultrasound-Former (Ours)	0.8183	0.7350	0.0355	0.8375	0.7601	0.0262	0.7686	0.6704	0.0596

we introduce axial attention for further saliency analysis of higher-level features. This consideration primarily addresses the complexity of ultrasound images, requiring increased focus on the object regions.

The axial attention is based on self-attention (Vaswani et al. 2017) which factorizes 2D attention into two 1D attention along height and width axes:

$$Attention_{row}(\cdot) = softmax\left(\frac{Q_{row}K_{row}^{T}}{\sqrt{d_k}}\right)V_{row}, \quad (6)$$

$$Attention_{col}(\cdot) = softmax\left(\frac{Q_{col}K_{col}^{T}}{\sqrt{d_k}}\right)V_{col}, \quad (7)$$

$$F_{Axial} = Attention_{col}(Attention_{row}(v)). \quad (8)$$

The reverse attention weight $W_{Reverse}$ is de-facto for salient object detection in the computer vision community (Fan et al. 2020; Kim, Lee, and Kim 2021), and can be formulated as:

$$W_{Reverse} = \Theta\left(\sigma\left(Up\left(S_{global}\right)\right)\right), \quad (9)$$

where $P(\cdot)$ denotes an up-sampling operation, $\sigma(\cdot)$ is the Sigmoid function, and $\Theta(\cdot)$ is a reverse operation subtracting the input from matrix $\mathbf{E}$, in which all the elements are 1. It is worth noting that the erasing strategy driven by reverse attention can eventually refine the imprecise and coarse estimation into an accurate and complete prediction map. Finally, we obtain the output boundary refinement features F_{BFM} by multiplying the axial attention output feature F_{Axial} by a reverse attention weight $W_{Reverse}$, as below:

$$F_{BFM} = W_{Reverse} \times F_{Axial}. \quad (10)$$

Experiments

Experimental Settings

Datasets and Evaluation protocols. We conduct experiments on the BUSI dataset (Al-Dhabyani et al. 2020). The dataset contains 780 images acquired by two types of ultrasound equipment (LOGIQ E9 ultrasound and LOGIQ E9 Agile ultrasound system) in the Baheya Hospital. The average image size of these images is 500×500 pixels. For quantitative comparison, we report three widely-used metrics including the mean Dice coefficient (mDice), mean Intersection over Union (mIoU), and mean absolute error (MAE). mDice and mIoU focus on the internal consistency of objects, while MAE represents the average value of the absolute error between the prediction and ground truth.

Implementation Details. We utilize a pre-trained PVT (Wang et al. 2021) model on ImageNet (Deng et al. 2009) as the backbone and conduct end-to-end training employing the AdamW optimizer (Loshchilov and Hutter 2017). The initial learning rate is set to 1e-4 and the weight decay is adjusted to 1e-4 too. Further, we resize the input images to 352×352 with a mini-batch size of 8 for 100 epochs. Given the diverse scales of objects in medical imaging, a multi-scale training is adopted following previous work (Dong et al. 2021). To ensure a fair comparison, the evaluation results of the comparative models within this paper are derived using the officially provided open-source code. All experiments are carried out utilizing PyTorch (Paszke et al. 2019) on a singular NVIDIA GeForce RTX 4070 GPU boasting 12 GB of memory.

Comparison with State-of-the-Art Methods

The quantitative evaluation results on ultrasound images are presented in Table 2. For breast lesions with indistinct boundaries, Ultrasound-Former achieves consistent improvements against baseline models under comparable mDice, mIoU and MAE. The visualization of our method and comparative methods are shown in Figure 5. Our method has the best performance in segmenting lesion. For instance, in the case of malignant tumor (first row, with serrated or lobulated boundary), other methods exhibit significant instances of false negatives, while our method addresses this issue well. Similarly, under blurred boundaries (second row) and speckle noise (third row) conditions, Ultrasound-Former exhibited no issues of either missed detections or false positives.

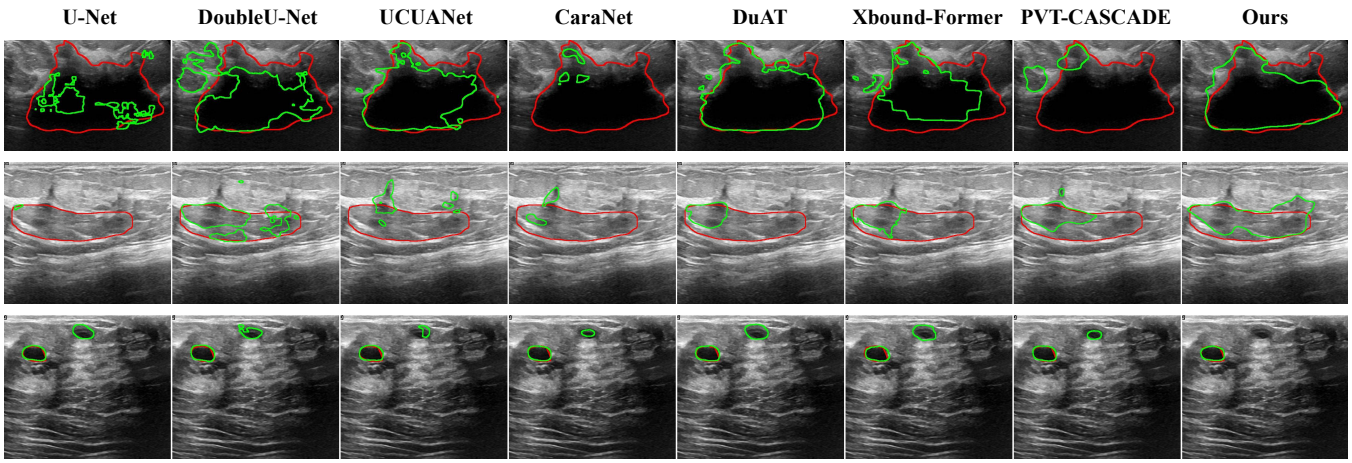

Figure 5: Qualitative comparison of different methods on BUSI (Al-Dhabyani et al. 2020). The red curve is the ground-truth boundary. The green curve is the segmentation results of our method.

Ablation Study

Effectiveness of Different Network Components. In Table 3, we employ a Transformer-based encoder combined with a partial decoder as our baseline. Note that our partial decoder is only deployed on the high-level features, which achieve an mDice score of 78.79%. This not only showcases the effectiveness of Transformer encoder but also highlights a natural advantage over conventional CNN-based methods. We further investigate the contribution of the Noise Suppression Module. We observe that adding NSM improves the baseline performance, increasing the mDice score from 78.79% to 80.31%. These improvements suggest that introducing NSM component can enable our model to enhance the quality of global maps. We verify the performance enhancement after integrating Boundary Refinement Module. A noticeable improvement of 1.93% in mDice score compared to the baseline is observed. This substantiates that BRM enables our model to accurately distinguish breast tumors. Finally, by simultaneously integrating the two primary components, we achieved a performance boost of 3.04%. This indicates that the fusion of high-quality coarse-grained information with refined boundary recovery is crucial and indispensable for localizing breast lesions.

Quantitative comparison of variants of NSM. As reported in Table 4, when showcasing variations of the NSM module, we aimed to demonstrate the impact of frequency information on speckle noise suppression. It's noteworthy that retaining only the low-pass filter within the NSM resulted in exceptional performance, surpassing an mDice score of 81%. However, introducing high-frequency information led to performance degradation, indicating that the cumulative noise error within the high-frequency data impaired the model's performance. Additionally, leveraging Gaussian filtering on top of the high-frequency filter effectively mitigated noise errors, preserving valuable high-frequency information and contributing to a 0.76% performance boost.

Table 3: Ablation study on the effectiveness of different components.

PD	NSM	RFM	mDice(%)	mIoU(%)
✓			78.79	70.16
✓	✓		80.31(+1.52)	71.49(+1.33)
✓		✓	80.72(+1.93)	72.10(+1.94)
✓	✓	✓	81.83(+3.04)	73.50(+3.34)

Table 4: Quantitative comparison of variants of NSM.

LPF	HPF(w/o denoise)	HPF	mDice(%)	mIoU(%)
✓			81.07	72.34
✓	✓		80.31(-0.76)	71.49(-0.85)
✓		✓	81.83(+0.76)	73.50(+1.16)

Conclusion

In this paper, we present an efficient Transformer based method for breast lesions segmentation in ultrasound images, named Ultrasound-Former, which utilizes a pyramid vision transformer backbone as the encoder to explicitly extract more powerful and robust features. Extensive experiments show that Ultrasound-Former consistently outperforms the current cutting-edge models on the BUSI dataset without any pre-/postprocessing. Specifically, we obtain the above-mention achievements by introducing two simple components, i.e., a noise suppression module (NSM) and a boundary refinement module (BRM), which effectively suppress the cumulative high-frequency internal errors caused by speckle noise and, from a practical perspective, optimize the boundary refinement process. We hope this research will stimulate more novel ideas for solving the breast lesion segmentation task in ultrasound images.

References

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in brief*.

Almajalid, R.; Shan, J.; Du, Y.; and Zhang, M. 2018. Development of a Deep-Learning-Based Method for Breast Ultrasound Image Segmentation. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*.

Chen, G.; Li, L.; Zhang, J.; and Dai, Y. 2022. Rethinking the Unpretentious U-net for Medical Ultrasound Image Segmentation.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dong, B.; Wang, P.; and Wang, F. 2023. Head-Free Lightweight Semantic Segmentation with Linear Transformer.

Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; and Shao, L. 2021. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*.

Hu, Y.; Guo, Y.; Wang, Y.; Yu, J.; Li, J.; Zhou, S.; and Chang, C. 2019. Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model. *Medical Physics*.

Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; and Wu, J. 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP*.

Jha, D.; Riegler, M. A.; Johansen, D.; Halvorsen, P.; and Johansen, H. D. 2020. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *IEEE International Symposium on Computer-Based Medical Systems*.

Kim, T.; Lee, H.; and Kim, D. 2021. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *ACM MM*.

Lee, H.; Park, J.; and Hwang, J. Y. 2020. Channel Attention Module with Multi-scale Grid Average Pooling for Breast Cancer Segmentation in an Ultrasound Image. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lou, A.; Guan, S.; Ko, H.; and Loew, M. H. 2022. CaraNet: Context axial reverse attention network for segmentation of small medical objects. In *Medical Imaging 2022: Image Processing*.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*.

Qi, W.; Wu, H.; and Chan, S. 2023. MDF-Net: A Multi-scale Dynamic Fusion Network for Breast Tumor Segmentation of Ultrasound Images. *TIP*.

Rahman, M. M.; and Marculescu, R. 2023. Medical Image Segmentation via Cascaded Attention Decoding. In *WACV*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Tang, F.; Huang, Q.; Wang, J.; Hou, X.; Su, J.; and Liu, J. 2022. DuAT: Dual-aggregation transformer network for medical image segmentation. *arXiv preprint arXiv:2212.11677*.

Valanarasu, J. M. J.; and Patel, V. M. 2022. Unext: Mlp-based rapid medical image segmentation network. In *MICCAI*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Chen, F.; Ma, Y.; Wang, L.; Fei, Z.; Shuai, J.; Tang, X.; Zhou, Q.; and Qin, J. 2023. XBound-Former: Toward cross-scale boundary modeling in Transformers. *TMI*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*.

Wei, J.; Hu, Y.; Zhang, R.; Li, Z.; Zhou, S. K.; and Cui, S. 2021. Shallow attention network for polyp segmentation. In *MICCAI*.

Yan, Y.; Liu, Y.; Wu, Y.; Zhang, H.; Zhang, Y.; and Meng, L. 2022. Accurate segmentation of breast tumors using AE U-net with HDC model in ultrasound images. *Biomedical Signal Processing and Control*, 103299.

Zhang, Y.; Liu, H.; and Hu, Q. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In *MICCAI*.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*.