FFR-UNet: Feature Filter-Refinement UNet for Medical Image Segmentation

Weixin Xu
xuweixin6666@gmail.com

Beihang University,
Beijing, China

Abstract

Medical image segmentation poses a significant challenge in the field of computer vision. Traditional approaches leverage Convolutional Neural Networks (CNNs) and Transformer-based methods to address the intricacies of medical image segmentation. However, inherent limitations persist: CNN-based methods often neglect long-range dependencies, and Transformer-based methods may overlook local context information. Moreover, in contrast to natural images, medical images present a distinct challenge wherein the foreground targets requiring segmentation are typically smaller, accompanied by a greater abundance of background information (considered as irrelevant information). This inherent characteristic often interferes with segmentation networks and leads to segmentation results that may lack the desired refinement. To overcome these deficiencies, we propose a novel Feature Filter Module (FFM) designed to discern between informative and non-informative features. These features seamlessly transition into our proposed Feature Refinement Module (FRM), assigning them distinct roles to establish a robust connection between the two input features. This strategy empowers our module to concurrently focus on both long-range dependencies and local context information by skillfully merging convolution operations with cross-attention mechanisms. Moreover, by integrating our proposed FFM and FRM into the encoder block of the UNet architecture, we introduce a novel framework named Feature Filter-Refinement UNet (FFR-UNet). Extensive experiments demonstrate the superiority of FFR-UNet, consistently achieving state-of-the-art (SOTA) performance compared to existing methods. Codes will be publicly available at https://github.com/xuweixinxxx/FFR-UNet.

1 Introduction

Accurate segmentation of medical images is crucial for pre-treatment diagnostics, treatment strategy formulation, and post-treatment evaluations across diverse medical conditions. The extensive adoption of convolutional neural networks (CNNs) in medical image segmentation tasks has gained momentum due to their efficacy in capturing representative features. The UNet model, proposed by Ronneberger [9], stands out as a pivotal architecture. Its distinctive encoder-decoder structure, enriched with skip connections, has demonstrated exceptional efficacy in segmenting medical images. Various adaptations, such as ResUNet [17], UNet++[18], and UNet3+[6], have been introduced, showcasing noteworthy performance in diverse medical image segmentation tasks while building upon the UNet foundation. Despite the success of these CNN-based methodologies, persistent challenges, notably in preserving

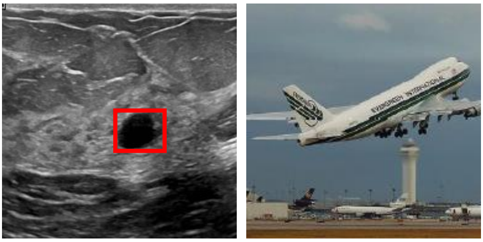

Figure 1: Example of the comparison between medical image and natural image.

long-range dependencies among pixels, stem from intrinsic limitations in the CNN architecture. Efforts have been made ([7], [11], [10]) to integrate attention modules into their frameworks, aiming to amplify feature maps and enhance pixel-level classification in medical images. However, effectively capturing extensive long-range dependencies remains a challenge, prompting the exploration of alternative methodologies.

The introduction of the transformer addresses the challenge of handling long-range dependencies effectively. Initially designed for machine translation in natural language processing (NLP)[14], the transformer has also found application in computer vision tasks through vision transformers (ViT)[3]. ViT excels in tasks like image classification and semantic segmentation by leveraging self-attention/cross-attention to learn correlations among input tokens, facilitating the capture of long-range dependencies. This involves dividing an image into non-overlapping patches fed into the transformer module with positional embeddings. Inspired by ViTs, some methods [2, 8] have been proposed for medical image segmentation, incorporating transformer blocks as middle layers. While these approaches have demonstrated satisfactory performance with transformer architectures, it's essential to note that the self-attention/cross-attention mechanisms inherent in transformers may limit their ability to learn local (contextual) relations among pixels, which are as important as global information while tackling medical image segmentation tasks. Moreover, as shown in Figure 1, compared with natural images, the lesion areas requiring segmentation in medical images often occupy a smaller proportion of the overall image and contain a higher amount of background. The presence of this surplus background, deemed as irrelevant information, can introduce interference to the model, ultimately yielding segmentation results that may lack the desired accuracy and completeness.

In this paper, to solve the above issues, we propose a novel Feature Filter Module (FFM) that filters out informative features and non-informative ones, to reduce the interference to the segmentation network. At the same time, to further refine the features and strengthen their interactions, we propose a Feature Refinement Module (FRM) to progressively suppress features in irrelevant background regions. Our main contributions are summarized as follows:

- We propose a novel **Feature Filter Module (FFM)**, aiming to enhance the network's ability to distinguish between informative and non-informative features to mitigate the impact of irrelevant features on the network and enhance the network's focus on crucial features.

- Following the FFM, we introduce an innovative **Feature Refinement Module**. This module integrates the convolutional operation and cross-attention mechanism to not only refine features but also enhance interactions among features from the preceding layer. The distinctive combination of these two operations allows us to focus on long-range dependencies and local relations concurrently. This dual-focus approach sets our FRM apart, enriching the model's capacity to capture intricate dependencies across different spatial scales.

- By incorporating the FFM and FRM into the encoder blocks of the UNet, we introduce a new framework dubbed **Feature Filter-Refinement UNet (FFR-UNet)**. We evaluate our proposed FFR-UNet on three widely used public benchmarks for medical image segmentation tasks. Extensive experiments demonstrate the proposed FFR-UNet can achieve state-of-the-art (SOTA) performance.

2 Method

2.1 Feature Filter Module

To effectively address the challenge of irrelevant features impacting the network and to elevate the precision of segmentation results, a meticulous distinction between informative and non-informative features is essential. In pursuit of this goal, we introduce the **Feature Filter Module (FFM)**, as illustrated in Figure 2. The FFM is meticulously crafted to discriminate between informative and non-informative features in a multi-step process, adding a layer of sophistication to the segmentation framework.

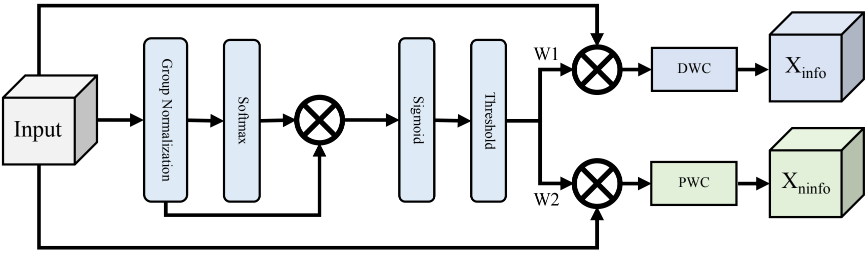

Figure 2: The framework of our proposed Feature Filter Module (FFM)

In detail, our proposed FFM encompasses the following steps: Firstly, a softmax operation, coupled with scaling factors in the Group Normalization (GN) [15] layers, is employed to assess the informative content of distinct feature maps, denoted as W_{info}, as calculated by Eq.(1). Subsequent to this evaluation, the weight values of feature maps, W_{info}, are normalized to the range (0, 1) through the sigmoid function, with a threshold of 0.5 applied to gate these values, as mathematically expressed in Equation (2). This step yields informative weights, W_1, by assigning weights above the threshold to 1. Simultaneously, non-informative weights, denoted as W_2, are obtained by assigning weights below the threshold to 0. Following this weight discrimination, element-wise multiplication is performed on the two sets of weighted outputs, W_1 and W_2, with the input feature X. This process culminates in the acquisition of informative features, denoted as X_{info}, through the application of depth-wise convolution (DWC) on the weighted X. Additionally, non-informative features, denoted as X_{ninfo}, are obtained by applying point-wise convolution (PWC) on the weighted X. In this manner, the comprehensive feature filtering process is effectively executed, thereby enhancing the model's ability to discern and prioritize informative features over non-informative ones.

$$W_{info} = GN(X) \times Softmax(GN(X)) \tag{1}$$

$$W = Threshold(Sigmoid(W_{info})) \tag{2}$$

2.2 Feature Refinement Module

Although non-informative features will introduce interference to the network, the rich structural information inherent is also crucial in medical images for segmentation tasks. Therefore, a direct and indiscriminate discarding of non-informative features is not advisable. This rationale underscores the introduction of the proposed **Feature Refinement Module (FRM)** following the feature filtering process, aiming to refine the features further and enhance their interactions.

After analyzing convolution and cross-attention mechanisms, it's clear that each has distinct strengths and weaknesses. Convolution excels in local information processing but may overlook global context and long-range dependencies. To address this, we integrate the cross-attention mechanism into our Feature Refine Module (FRM), combining the strengths of both methods. This strategic fusion enables a comprehensive capture of intricate dependencies among extracted features, significantly enhancing our ability to refine embedded information by aggregating feature maps.

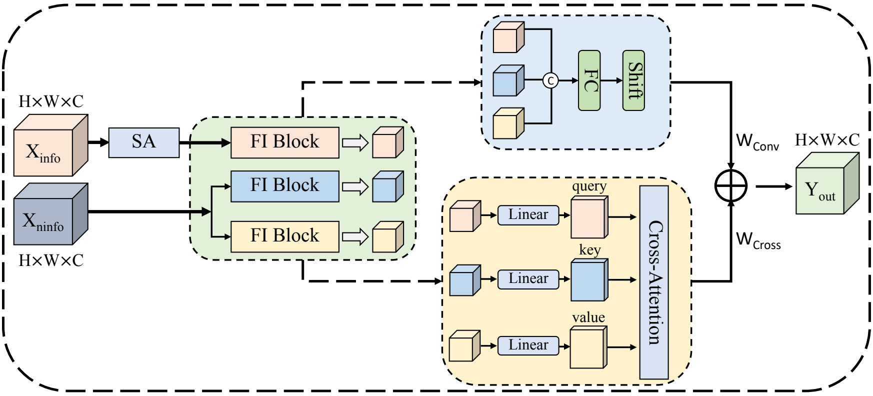

Figure 3: The framework of our proposed Feature Refinement Module (FRM). SA represents the Self-Attention module.

As shown in Figure 3, the FRM receives two inputs, X$_{info}$ and X$_{ninfo}$. Specifically, our Feature Refine Module (FRM) comprises the following steps: Firstly, we employ a self-attention to initialize the informative features, X$_{info}$, to get the X$^{I}_{info}$. Then, two input features undergo processing by three **Feature Initialize (FI)** Blocks, each incorporating a 1×1 convolution and a ConvMixer block [12]. During this step, X$^{I}_{info}$ is processed by one of these blocks, while X$ninfo$ undergoes processing by two. Subsequently, all processed results are fed into two distinct pathways: the convolution path and the cross-attention path. For the convolution path (eq. 3 and 4, three features are concatenated first and then fed into a fully connected layer. Subsequent to this step, a shift operation is judiciously utilized to yield the ultimate output feature, denoted as Y_{conv}. Within the cross-attention path, three input features from the previous projected layer are meticulously fed into. After linear embedding layers, these features are thoughtfully rendered as query (originating from X_{info}), key, and value, respectively. Consequently, these extracted features are introduced into the cross-attention mechanism, culminating in the derivation of the definitive outcome, dubbed as Y_{cross}. The final output is determined through the integration of these two paths, achieved by the addition of Y_{conv} and Y_{cross}, their respective contributions modulated by two learnable scalars (W_{Conv} and W_{Cross}), as delineated in eq. (5).

$$Y = Concat(FI_1(X_{info}^{I}), FI_2(X_{ninfo}), FI_3(X_{ninfo})) \tag{3}$$

$$Y_{conv} = Shift(FC(Y)) \tag{4}$$

$$Y_{out} = W_{Conv} \times Y_{conv} + W_{Cross} \times Y_{cross} \tag{5}$$

2.3 Overall Architecture

The architectural framework of our Feature Filter-Refinement UNet (FFR-UNet) is illustrated in Figure 4. Adhering to the foundational tenets of the UNet framework [9], our design ethos prioritizes the preservation of the encoder-decoder structure. This deliberate choice is underpinned by the strategic integration of skip connections, a measure intended to harness their inherent advantages. The incorporation of skip connections not only facilitates the extraction of richer feature representations but also serves as a robust defense mechanism against network degradation and the potential pitfalls associated with overfitting.

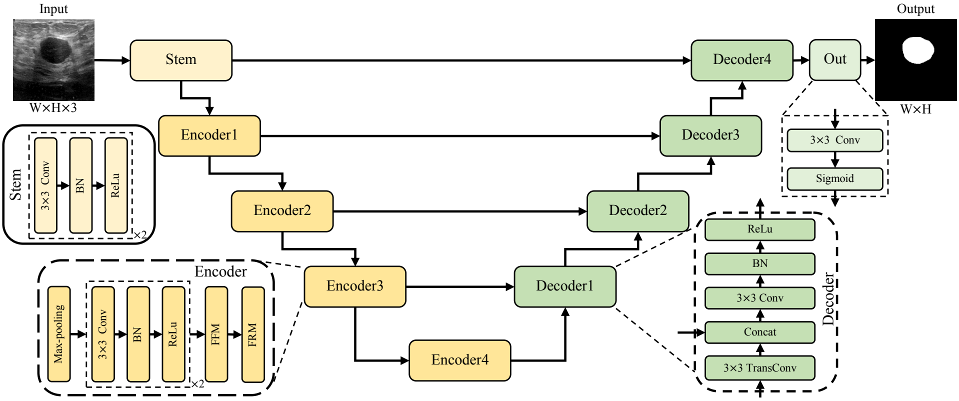

Figure 4: The framework of our proposed Feature Filter-Refinement UNet (FFR-UNet). The Concat represents the channel-wise stack. We use 3×3 kernel with stride as 2 for Transpose Convolution (Transconv) in the decoder layers.

Within the confines of our architectural framework, we recognize that features derived from the convolutional layers encompass both informative and non-informative components. Being aware of the potential interference introduced by non-informative features in subsequent layers, addressing this concern is paramount to ensure the final segmentation results meet stringent quality standards. To this end, we introduce the Feature Filter Module (FFM) into the encoder layers, strategically positioned to discriminate between informative and non-informative features. Elevating this objective further, the Feature Refinement Module (FRM) is introduced. By synergistically integrating convolution and cross-attention mechanisms, the FRM is designed to refine features, fostering heightened interaction among distinct feature components. This nuanced approach aims to strike a delicate balance between preserving long-range dependencies and capturing local context information effectively, thereby enhancing the overall segmentation accuracy of the FFR-UNet.

3 Experiments

3.1 Datasets and Evaluation Metrics

3.1.1 Datasets

Three datasets are involved in our study. The BUSI dataset [1] comprises 780 breast ultrasound images, averaging 500 × 500 pixels, featuring normal, benign, and malignant cases of breast cancer with corresponding segmentations. Our focus on benign and malignant images (647 total) led to a balanced 7:1:2 split for training (453 images), validation (65 images), and testing (129 images). The BUSIS dataset [16] includes 562 images from women aged 26 to 78 years, randomly divided into training (394 images), validation (56 images), and test sets (112 images) following a balanced 7:1:2 ratio. The TN3k dataset [5],[4] consists of 3493 thyroid ultrasound images with high-quality nodule masks, distributed for training and validation (2879 images) and testing (614 images) as per the official split[5],[4].

3.2 Evaluation Metrics

We employ four commonly used metrics to perform a quantitative assessment of the performance of various segmentation models. These metrics include the Dice Similarity Coefficient (DSC), mean Intersection over Union (mIoU), Precision, and Recall. The calculation formulas are defined as follows:

$$IoU = \frac{TP}{TP+FN+FP} \tag{6}$$

$$DSC = \frac{2 \times TP}{(TP+FN)+(TP+FP)} \tag{7}$$

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

3.3 Implementation Details

Our experimental setup leveraged PyTorch 1.7.0, ensuring compatibility with contemporary deep learning frameworks. The training process was conducted on a singular NVIDIA RTX A6000 GPU endowed with 48GB of memory. For the optimization process, we employed the Adam optimizer with a learning rate annealing factor set at 0.2. The learning rate was initialized at 1e-4, and a weight decay of 5e-4 was applied to control overfitting. These hyperparameters were chosen to strike a balance between rapid convergence during training and preventing the model from overfitting to the training data.

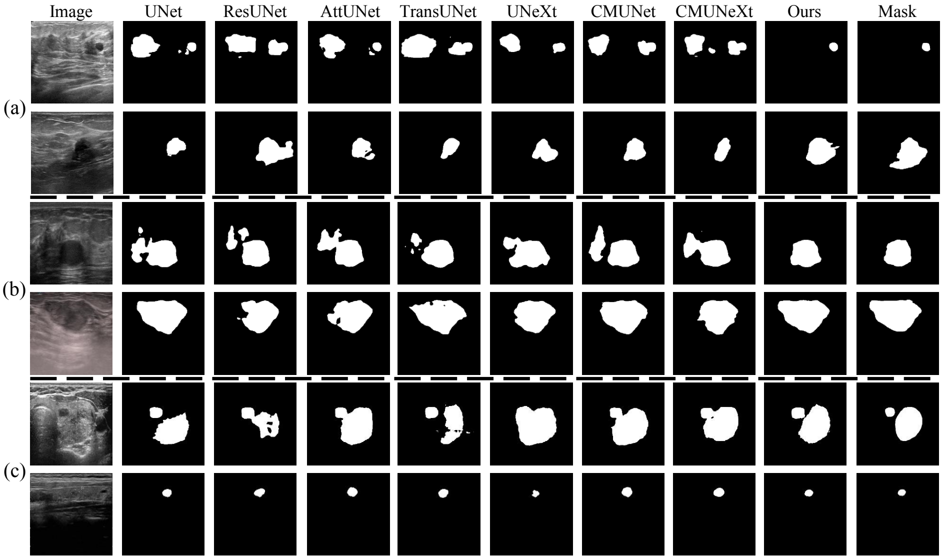

Figure 5: Visualized results on BUSI, BUSIS, and TN3K datasets. From left to right are the input image, results of UNet, ResUNet, AttUNet, TransUNet, UNeXt, CMUNet, CMUNeXt, our proposed method, and the ground truth mask, respectively. (a) represents the segmentation results of benign and malignant nodules from the BUSI dataset; (b) represents segmentation results on the BUSIS dataset; and (c) represents results on the TN3K dataset.

In terms of the loss function, we adopted a strategy aligned with prior works [10, 11]. Specifically, we combined the dice loss and binary cross-entropy, as defined in eq. (10). This combination aims to capitalize on the complementary strengths of both loss functions, fostering a more comprehensive optimization strategy. The dice loss is particularly useful in handling class imbalance, a common challenge in medical image segmentation tasks. By emphasizing the importance of rare or small anatomical structures, the dice loss helps the model achieve more accurate segmentations. On the other hand, the binary cross-entropy loss contributes to the overall stability and convergence of the training process, ensuring smooth optimization.

$$Loss = Dice(pred, gt) + 0.5 \times BCE(pred, gt) \tag{10}$$

3.3.1 Data Preprocessing

In addressing the constraints imposed by GPU memory limitations, our experimental setup employs a batch size of 8, and each model undergoes training for a duration of 100 epochs. To ensure uniformity and facilitate efficient processing, we conform to the image resizing strategy outlined in [11], where all images are resized to $256 \times 256 \times 3$. To counteract the potential challenges associated with insufficient data and mitigate the risk of overfitting, we incorporate data augmentation techniques into the training set. Specifically, we apply random rotation, flip, elastic transform, and light transforms to the images, each with a probability of 0.5. It is crucial to note that these augmentation operations are exclusively performed on the training set, preserving the integrity of the validation and test sets.

3.4 Results

In our comprehensive evaluation, we benchmark our proposed methods against seven state-of-the-art models, including UNet, ResUNet, AttUNet, TransUNet, UNeXt, CMUNet, and CMUNeXt. The detailed results obtained on the BUSI and BUSIS datasets are presented in Table 1. Noteworthy advancements in performance metrics are observed when compared with previous state-of-the-art (SOTA) methods. Specifically, for the BUSI dataset, our model showcases improvements of 0.98% and 0.68% in mIoU and DSC, respectively. Similarly, on the BUSIS dataset, our method demonstrates enhancements of 0.75% and 0.48% in mIoU and DSC, respectively. The superior performance of our proposed method is further validated on the TN3K dataset, where we achieve significant improvements of 1.81% and 1.30% in terms of mIoU and DSC, as illustrated in Table 2.

It is worth mentioning that the improvements achieved by our method are not just numerical, but also have practical implications. In medical image analysis, accurate segmentation is crucial for identifying anatomical structures, pathologies, and other relevant features. The enhanced performance of our method can lead to more precise and reliable segmentation results, which can, in turn, improve the accuracy of diagnoses and treatment plans. Moreover, our method's superiority is not limited to quantitative metrics but also extends to qualitative assessments. The segmentation maps produced by our method tend to be smoother and more contiguous, with fewer artifacts and misclassifications.

Methods	BUSI dataset (%)				BUSIS dataset (%)			
	DSC	mIoU	Precision	Recall	DSC	mIoU	Precision	Recall
UNet [9]	76.18	67.62	79.09	79.71	91.18	84.89	93.02	90.88
ResUNet [17]	77.27	68.45	79.20	80.36	91.26	85.09	93.18	91.15
AttUNet [7]	76.62	68.09	79.72	78.44	91.04	84.65	93.04	90.66
TransUNet [2]	71.94	61.88	78.57	73.57	89.97	82.69	90.09	91.53
UNeXt [13]	72.27	61.64	77.06	75.39	89.97	82.41	90.69	90.57
CMUNet [11]	79.95	71.68	84.13	81.45	91.43	85.28	92.86	91.58
CMUNeXt [10]	74.52	65.32	77.53	76.46	90.43	83.41	90.89	91.29
Our Model	**81.17**	**72.92**	**82.41**	**82.72**	**92.11**	**86.26**	**93.73**	**91.80**

Table 1: Comparision results on BUSI and BUSIS datasets.

Methods	TN3K dataset (%)			
	DSC	mIoU	Precision	Recall
UNet [9]	77.69	67.38	74.91	87.08
ResUNet [17]	76.76	66.67	73.93	86.18
AttUNet [7]	77.80	67.86	74.94	87.04
TransUNet [2]	71.65	60.03	69.37	83.13
UNeXt [13]	74.35	63.15	72.07	85.19
CMUNet [11]	79.86	70.15	78.35	85.19
CMUNeXt [10]	75.83	65.73	73.39	85.34
Our Model	**81.16**	**71.96**	**80.28**	**87.49**

Table 2: Comparision results on TN3K datasets.

To provide a visual representation of our method's efficacy, Figure 5 displays a selection of segmentation results. The results unequivocally highlight the superior performance of our proposed method in achieving more comprehensive nodule segmentation while effectively mitigating disturbances compared to competing methods. The images in Figure 5 showcase the consistent excellence of our model across diverse scenarios, including small and large nodules, as well as challenging nodules that closely resemble the background, making them difficult to differentiate. Furthermore, our model's segmentation results outperform other methods in terms of completeness, shape similarity to the ground truth (GT), and minimization of false positive regions. Notably, even in scenarios with multiple nodules within an image, our model adeptly segments them and successfully excludes interference from unrelated background regions. These visualized results provide compelling evidence of the superior performance and robustness of our proposed approach.

3.5 Ablation Study

The ablation study results, detailed in Table 3 and 4, provide insightful analysis into the efficacy of our proposed Feature Filter Module (FFM) and Feature Refinement Module (FRM) on medical image segmentation tasks. The experimental outcomes reveal compelling improvements in performance metrics, emphasizing the significance of our proposed modules.

For the BUSI and BUSIS datasets, the incorporation of our FFM, coupled with the replacement of FRM by direct summation or concatenation operations, yields notable enhancements. Specifically, we observe improvements of **(2.24%, 2.60%)** and **(0.26%, 0.28%)** in mIoU and DSC, respectively. Furthermore, the introduction of our proposed FRM contributes to additional performance gains, leading to a cumulative improvement of **(3.06%, 2.09%)** in mIoU and **(1.11%, 0.65%)** in DSC metrics. Extending our analysis to the TN3K dataset, the results underscore the effectiveness of FFM, demonstrating **0.89%** and **0.34%** improvement in mIoU and DSC metrics compared to the baseline model. However, it's noteworthy that the performance metrics achieved by methods leveraging our proposed FRM surpass those utilizing only FFM, with **3.69%** and **3.19%** higher mIoU and DSC metrics, respectively. These findings emphasize the pivotal role of our proposed FRM in further enhancing segmentation quality.

In conclusion, the comprehensive ablation study affirms the efficacy and indispensability of our proposed FFM and FRM modules in the context of medical image segmentation tasks.

Methods	BUSI dataset (%)				BUSIS dataset (%)			
	DSC	mIoU	Precision	Recall	DSC	mIoU	Precision	Recall
baseline	76.18	67.62	79.09	79.71	91.18	84.89	93.02	90.88
Ours w/o FRM	79.08	69.86	81.30	81.96	91.46	85.15	93.25	90.98
Our Model	**81.17**	**72.92**	**82.41**	**82.72**	**92.11**	**86.26**	**93.73**	**91.80**

Table 3: Ablation study results on BUSI and BUSIS datasets.

Methods	TN3K dataset (%)			
	DSC	mIoU	Precision	Recall
baseline	77.69	67.38	74.91	87.08
Ours w/o FRM	78.03	68.27	74.64	**87.61**
Our Model	**81.16**	**71.96**	**80.28**	**87.49**

Table 4: Ablation study results on TN3K datasets.

4 Conclusion

In this paper, we unveil the Feature Filter-Refinement UNet (FFR-UNet), an innovative framework meticulously crafted to not only elevate the precision of segmentation but also to augment the refinement of features in the domain of medical image segmentation. At the heart of our methodology lies the introduction of a groundbreaking Feature Filter Module (FFM), strategically deployed to process features extracted from the preceding layer. The FFM, with its inherent capability, adeptly discriminates between informative and non-informative features, paving the way for a more nuanced understanding of the underlying image data. Building upon the FFM, our contribution extends to the introduction of the Feature Refinement Module (FRM). The FRM, a carefully devised amalgamation of convolution operations and cross-attention mechanisms, emerges as a powerful tool for refining features. This module not only enhances feature precision but also ensures the incorporation of both long-range dependencies and local context information, thus presenting a holistic approach to feature refinement. The culmination of our efforts in the seamless integration of the FFM and FRM into the encoder layers of the UNet architecture gives rise to the specialized FFR-UNet tailored explicitly for medical image segmentation.

References

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

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 257–261, 2021. doi: 10.1109/ISBI48211.2021.9434087.

[5] Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Fei Chen, and Guanbin Li. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in Biology and Medicine*, 106389:1–12, 2022.

[6] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[7] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[8] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 267–276. Springer, 2021.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[10] Fenghe Tang, Jianrui Ding, Lingtao Wang, Chunping Ning, and S Kevin Zhou. Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. *arXiv preprint arXiv:2308.01239*, 2023.

[11] Fenghe Tang, Lingtao Wang, Chunping Ning, Min Xian, and Jianrui Ding. Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In *2023*

IEEE 20th International Symposium on Biomedical Imaging (ISBI), pages 1–5. IEEE, 2023.

[12] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

[13] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 23–33. Springer, 2022.

[14] 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.

[15] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[16] Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. Busis: a benchmark for breast ultrasound image segmentation. In *Healthcare*, volume 10, page 729. MDPI, 2022.

[17] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[18] 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*, 39(6):1856–1867, 2019.