

Received 14 March 2024, accepted 9 May 2024, date of publication 13 May 2024, date of current version 20 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3400319

RESEARCH ARTICLE

MFI-Net: Multi-Level Feature Integration Network With SE-Res2Conv Encoder for Jaw Cyst Segmentation

HUIXIA ZHENG¹, XIAOLIANG JIANG², XU XU¹, AND XIAOKANG DING²

¹Department of Stomatology, Quzhou People's Hospital, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou 324000, China

²College of Mechanical Engineering, Quzhou University, Quzhou 324000, China

Corresponding authors: Xiaoliang Jiang (jxl_swjtu@163.com) and Xu Xu (cook10102@163.com)


This work was supported in part by the National Natural Science Foundation of China under Grant 62102227; in part by Zhejiang Basic Public Welfare Research Project under Grant LZY24E050001, Grant LZY24E060001, and Grant ZCLTGS24E0601; and in part by the Science and Technology Major Projects of Quzhou under Grant 2023K221, Grant 2023K211, Grant 2023K140, and Grant 2021029.

ABSTRACT Accurate segmentation of organs and lesions from medical images holds paramount importance in aiding physicians with diagnosis and monitor diseases. At present, the widespread application of deep-learning in medical image segmentation is primarily attributed to its exceptional feature extraction capability. Nonetheless, due to blurred target boundary, wide range of changes and chaotic background, the segmentation of medical images is still faced with great challenges. To address these issues, we present a multi-level feature integration network (MFI-Net) with SE-Res2Conv encoder for jaw cyst segmentation. Specifically, we replace the original convolution operation with SE-Res2Conv to better maintain model's capacity for extracting features across multiple scales. Then, a novel context extractor module including multi-scale pooling block (MPB) and position attention module (PAM), which aims to generate more discriminative features. Finally, a multi-level feature integration block (MFIB) is implemented within the decoder to efficiently integrate low-level detail features with high-level semantic features. Numerous experiments were conducted on both the original and augmented datasets of jaw cyst to demonstrate the advantages of MFI-Net, with results consistently superior to all competitors. The Dice, IoU and Jaccard values of our method reached 93.06%, 93.47%, 87.06% in the original database and 91.25%, 91.94%, 84.06% in the augmented database. Furthermore, the computational efficiency of MFI-Net is impressive, with a speed of 106.21 FPS and 110.28 FPS at the input size of $3 \times 256 \times 256$ on a NVIDIA RTX6000 graphics card.

INDEX TERMS Image segmentation, jaw cyst, deep learning, SE-Res2Conv, multi-level feature integration.

I. INTRODUCTION

As a diverse benign pathological entity, jaw cyst can be manifested in various forms, including rhizoid cyst, dental cyst and odontogenic keratocyst, each of which has its own unique morphological and clinical characteristics. In addition, jaw cysts are often a precursor to complications that can lead to tooth impaction, bone resorption, and the potential to worsen a patient's condition if left unaddressed. In the ever-evolving field of health, medical imaging has become a cornerstone for

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva .

clinicians and diagnostics to unravel the mysteries hidden in the human body. Within this vast expanse, diagnostic radiology emerged as a key discipline, harnessing the alchemical power of high-resolution images to reveal complex details of anatomy and pathology. Traditionally, the responsibility for identifying and delineating these cystic structures has rested solely on the oral and maxillofacial surgeon. Relying on their professional experience and clinical acumen, these practitioners have meticulously traced the boundaries of jaw cysts through manual segmentation. However, this approach comes with inherent limitations, and it requires a significant investment of time and effort, resources that may be scarce

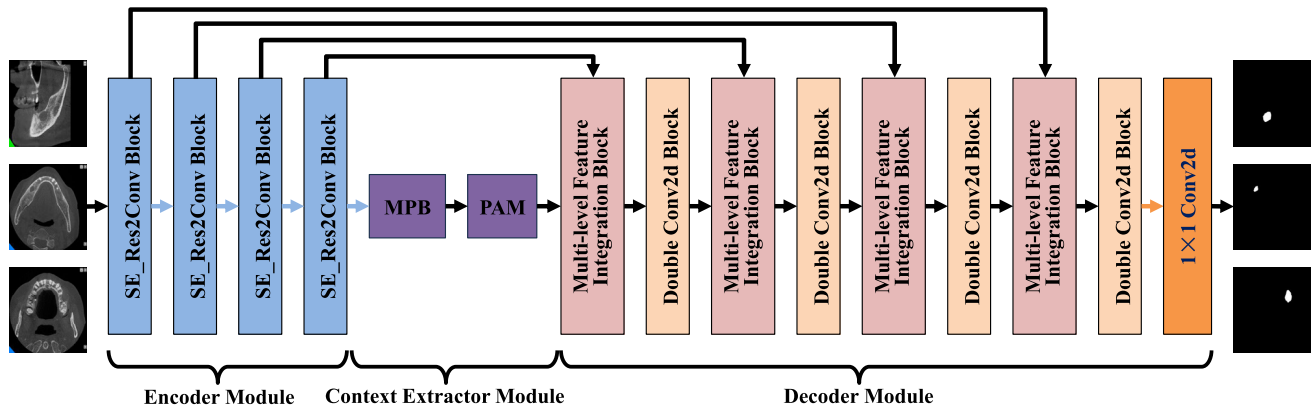


FIGURE 1. Network architecture of our MFI-Net.

in a fast-paced clinical setting. Moreover, the possibility of inter-observer variability and diagnostic inconsistencies are introduced. To address these multifaceted challenges, researchers can take advantage of the latest technologies in the field of medical image analysis, especially machine learning and deep learning. By harnessing the computational power of these technologies, a large part of the diagnostic workload can be reduced and the efficiency and accuracy of jaw cyst diagnosis can be significantly improved.

Early image segmentation techniques, such as statistical shape [1], [2], level set [3], [4], fuzzy clustering [5], [6]. Each approach has its own unique set of parameters that can be fine-tuned to meet the specific needs of different medical image scenarios. Although they play an important role in the field of medical images, these algorithms still have certain limitations when dealing with complex data sets. In addition, the segmentation performance can also be significantly affected by external factors, including image acquisition quality, uneven lighting, and complex organized backgrounds. In response to the above problems, various methods based on deep learning [7], [8], [9], [10], [11], [12], [13] have become the mainstream in the field of medical image segmentation due to their powerful feature learning, end-to-end training, adaptability and generalization. At present, many researchers focus on fully convolutional networks (FCN) [14] and U-shaped structures [15], [16], [17], [18], [19], which usually employ encoder-decoder frameworks and thus perform well in simulating local features of images. However, due to the limited receptive field of convolution operations, these methods are often difficult to capture the global dependence of features, which is particularly important in semantic segmentation.

Recently, more variant networks have been proposed, mainly including recurrent neural networks [20], [21], multi-scale features [22], [23], residual connections [24], [25], attention mechanisms [26], [27]. Among them, Zhang et al. [28] introduced an encoder-decoder architecture integrating multi-scale contextual information. Initially, the encoder layer incorporates iterative input of the probabilistic mapping derived from the preceding classifier, facilitating the

fusion of high-order shape context and low-order appearance features across multiple scales. Subsequently, dense connectivity was employed to aggregate feature maps from encoders and decoders operating at various scales. Li et al. [29] presented an innovative approach to ophthalmic OCT segmentation by integrating recursive residual networks with attention mechanisms. Firstly, the recursive residual convolutional network was introduced to address issues such as image drilling and rapid degradation. Moreover, the attention mechanism was incorporated to enhance the utilization of global image information. Mubashar et al. [30] proposed R2U++, a novel method for medical image segmentation. Departing from conventional methods, this architecture substitutes the standard convolutional backbone with a deeper recursive residual convolutional block, which is more effective extraction of key features for segmentation. Additionally, the integration of a dense jump path serves to mitigate the semantic gap between the encoder and decoder modules. Xiong et al. [31] utilized compression-excitation and attention modules to construct a helically closed pathway, and presented a novel U-Net (SEA-Net) for precise small target segmentation.

Inspired by the above methods, this study presents a multi-level feature integration network with SE-Res2Conv encoder for jaw cyst segmentation. The experiment results on both the original and augmented datasets of jaw cyst show that the proposed MFI-Net network achieves significant segmentation performance. Our major contributions can be drawn as: 1) SE-Res2Conv was introduced as the encoder, which can better maintain model's capacity in extracting features across various scales effectively. 2) Both the MPB and PAM were incorporated into the context extractor module to generate more discriminative features. 3) A multi-level feature integration block was designed to integrate intricate low-level detail features with broader high-level semantic features.

II. METHODS

In this section, a multi-level feature integration approach is proposed for the precise segmentation of jaw cysts. Our

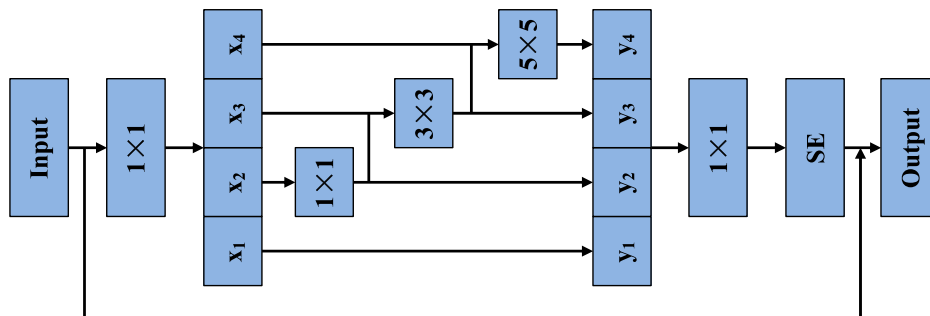


FIGURE 2. Structure of SE-Res2Conv block.

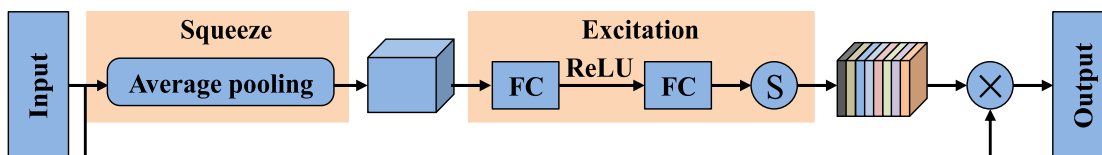


FIGURE 3. Structure of SE block.

method consists of three pivotal parts: encoder module, context extractor module and decoder module, as delineated in Figure 1. Detailed introduction will be given in the following subsequent chapters.

A. FRAMEWORK OF MFI-Net

To fully utilize the available data information while reducing the substantial computational resources typically required, we chose the encoder-decoder solution provided by U-Net architecture. In our approach, we carefully build each component to meet its corresponding challenges and optimize performance. Specifically, for the encoder module, we apply the SE-Res2Conv block as a replacement for the conventional convolution operation. This substitution not only avoids the common problems of disappearing gradients and exploding gradients, but also ensures efficient feature extraction without incurring additional computational overhead. In addition, the squeeze-and-excitation (SE) [32], [33] block is introduced to enable network to dynamically enhance the response of useful feature channels. Next, moving to the context extractor, we develop a newly module that consists of the MPB and the PAM. With this combination, it performs well at learning complex semantic contexts and can provide more nuanced feature maps. Finally, in the decoder stage, a multi-level feature integration block is designed to yield better fusion performance of low-level detail features and high-level semantic features. In the following subsections, we will provide a comprehensive introduction to each of these components to clarify their performance contributions.

B. SE-Res2Conv BLOCK

As depicted the architecture of SE-Res2Conv in Figure 2, an input feature map is splits into four distinct groups

represented as $\{x_1, x_2, x_3, x_4\}$ based on their channel attributes after the initial 1×1 convolution. Within this framework, the first subset x_1 is directly transmitted to the corresponding output y_1 without any additional processing, preserving essential features. The second subset x_2 is convolved by 1×1 kernel and further divided into two distinct pathways, one continues its propagation towards the designated output y_2 while the other diverges to merge with the subsequent segment x_3 , so that the third subset can obtain contextual information from previous layer. Similarly, the derivation of y_3 and y_4 is similar to that of y_2 , except that the third and fourth subsets utilize 3×3 and 5×5 convolutional kernel. This diverse convolutional kernel architecture enables the extraction of features at multiple spatial scales, enriching the model’s understanding of the input data’s intricacies. Then, the subsets $\{y_1, y_2, y_3, y_4\}$ are merged and a 1×1 convolution layer is applied to refine and consolidate the extracted features. After that, we introduce SE block to dynamically enhance the response of useful feature channels, and its structure is shown in Figure 3. In addition, in order to prevent over-fitting and ensure the robustness and generalization of the model, residual connection is employed to mitigate the risk of feature degradation during training. In summation, the modules built in the SE-Res2Conv architecture produce a fine output segmentation prediction image, characterized by enhanced feature representation and context understanding, thus emphasizing the effectiveness of the architecture in complex data processing tasks.

C. MULTI-SCALE POOLING BLOCK

In the context of jaw cyst images, the challenge of lesion size change is a major obstacle in the segmentation process. To solve this challenge, we present a multi-scale pooling

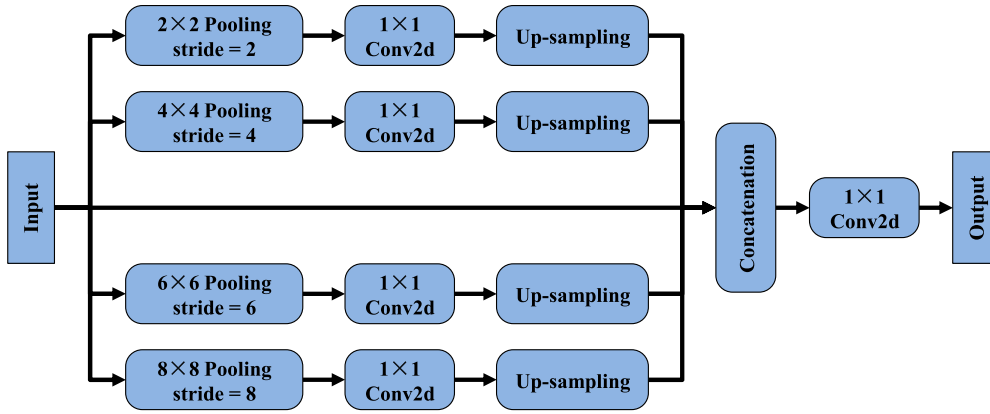


FIGURE 4. Structure of multi-scale pooling block.

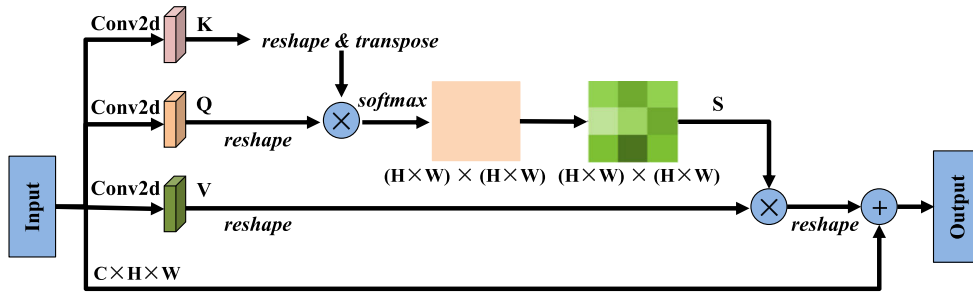


FIGURE 5. Structure of position attention module.

block that relies on multiple effective fields of view to detect targets of different sizes, as shown in Figure 4. Specifically, the MPB strategically routes the incoming input image features and passes them along five distinct paths. Among these paths, one remains unaltered, directly passing the feature map without any operation. However, the other four paths use maximum pooling operations to reduce the resolution to 1/2, 1/4, 1/6, and 1/8 of the original, thereby acquiring different receptive field sizes. Following this phase, a 1×1 convolutional layer is applied to each path to extract and assimilate multi-scale contextual information while simultaneously mitigating computational overhead by reducing the dimensionality of weights. Subsequently, the feature maps from the convolutional layer are up-sampled using bilinear interpolation to restore them to the original dimensions and retain important spatial information. Finally, the features obtained from all five paths are connected and passed through the final 1×1 convolutional layer to generate the final output features. In this study, the step size and the kernel dimensions are set to 2, 4, 6, 8, MPB can obtain the best segmentation effect.

D. POSITION ATTENTION MODULE

The role of the position attention module is to promote accurate and context-aware attention mechanisms within the network framework. For an input feature map $I \in \mathbb{R}^{C \times H \times W}$, the variables C, H, W representing the channel, height, and

width dimensions. As shown in Figure 5, the module has multiple branches, each designed to extract and manipulate essential information from the input. In the first and second branches, the input I is passed through the convolution operation to obtain two distinct feature mappings, namely K and Q . Following their derivation, K and Q are subjected to a multiplicative interaction, and then the softmax function is performed on the resultant matrix to obtain the position attention map $S^p \in \mathbb{R}^{(H \times W) \times (H \times W)}$:

$$S_{ij}^p = \frac{\exp(K_i \cdot Q_j)}{\sum_{i,j=1}^{H \times W} \exp(K_i \cdot Q_j)} \quad (1)$$

where S_{ij}^p denotes the influence of the position i th on the position j th.

Subsequently, the feature map I is processed with a convolution layer in the third branch, resulting in $V \in \mathbb{R}^{C \times (H \times W)}$. Similar to the other branches, the transformed feature map V is subjected to multiplication by a permuted version of the position attention map S^p . Furthermore, to augment the significance of this integration, the output is scaled by a factor α and adding it to I , then the final output O is formulated:

$$O_j = \alpha \sum_{i=1}^{H \times W} (S_{ij}^p V_i + I_j) \quad (2)$$

It is obvious from Equation (2) that the output map O aggregates the features not only from its own position but also

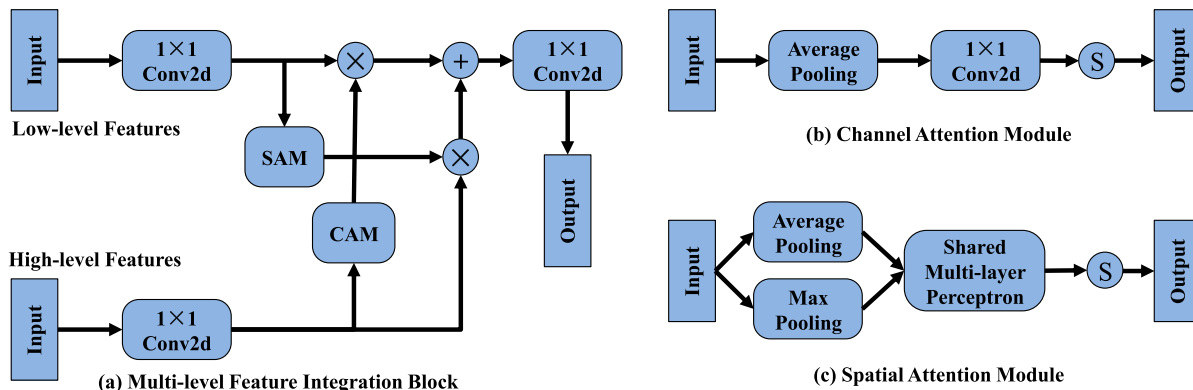


FIGURE 6. Structure of multi-level feature integration block.

incorporates the contributions of other locations, so that it can expertly identify and prioritize relevant location information.

E. MULTI-LEVEL FEATURE INTEGRATION BLOCK

In the U-shaped framework, the decoder is designed for the process of up-sampling the feature map generated by the encoder and gradually restores the resolution. In addition, it also undertakes a critical feature fusion operation. In traditional deep learning architectures, it is difficult to capture global information from encoders through operations like deconvolution, de-pooling, and bilinear up-sampling. Previous studies have shown that the lower-level features have higher resolution and contain finer location and detail information, while the higher-level features exhibit stronger semantic attributes. To improve the effect of algorithm, we combine the low-level detail features with high-level semantic features, and propose a multi-level feature integration block within the decoder, as shown in Figure 6. The multi-level feature integration block is composed of two modules: channel attention module (CAM) and spatial attention module (SAM). Firstly, in the initial stage, a 1×1 convolution is applied to both high-level and low-level features. Subsequently, we use the CAM go through an average-pooling, 1×1 convolution, and the sigmoid function for the high-level features, and then multiplies it with the low-level features. While in the SAM, the low-level features are averaged and maximally process to yield two-channel features. These components are then meticulously channeled into a shared multilayer perceptron architecture, where the sigmoid function calculates the spatial attention feature map, subsequently multiplied by up-sampled high-level features. Finally, the above two features undergo addition, followed by a 1×1 convolution to produce the final feature map. By integrating semantic concepts and spatial details, the multi-level feature integration block can enable networks to extract subtle differences from complex data with greater efficiency.

F. LOSS FUNCTION

Given the constraints of a small dataset, coupled with variations in lesion sizes and disparities in foreground and background distributions, the likelihood of encountering

class imbalance issues is substantially heightened. Unlike Cross-entropy loss [34], Dice loss [35] is widely favored for its ability to handle class imbalance and its effectiveness in producing accurate segmentation results even with limited annotated data. Therefore, Dice loss is adopted in this paper, and its calculation formula is as follow:

$$L_{Dice} = 1 - \frac{2 \sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N y_i^2 + \sum_{i=1}^N \hat{y}_i^2} \tag{3}$$

where N is the number of pixels, \hat{y}_i and y_i represent the predicted segmentation masks and actual label, respectively.

III. RESULTS

A. DATASET

In this section, we meticulously assess the performance of our MFI-Net on both the original and augmented datasets of jaw cyst. The details of these datasets are as follows.

Original Dataset: We cooperated with the stomatology department of Quzhou People’s Hospital to collect a dataset of jaw cyst images. The images were primarily obtained using the oral-maxillofacial conical beam computed tomography device, and the dataset itself contained a large collection of 1535 images, of which 922 images were allocated for training, 307 images were set aside for validation, and 306 images were reserved for rigorous testing to assess the generalizability and performance of our algorithms. Recognizing the importance of standardization in data preprocessing, each image is subjected to a rigorous adjustment procedure with a size of 256×256 pixels.

Augmented Dataset: Different from conventional image datasets, the collection of medical images is more difficult, and the accuracy and reliability of the marks are strictly required. However, insufficient training data is easy to over-fitting in the training stage, which leads to the reduction of algorithm accuracy. To mitigate the risks associated with insufficient training data and over-fitting, we employ data augmentation strategy (including rotation, scaling shifts, translation, clipping), which is designed to diversify the training data while maintaining its semantic integrity. In the

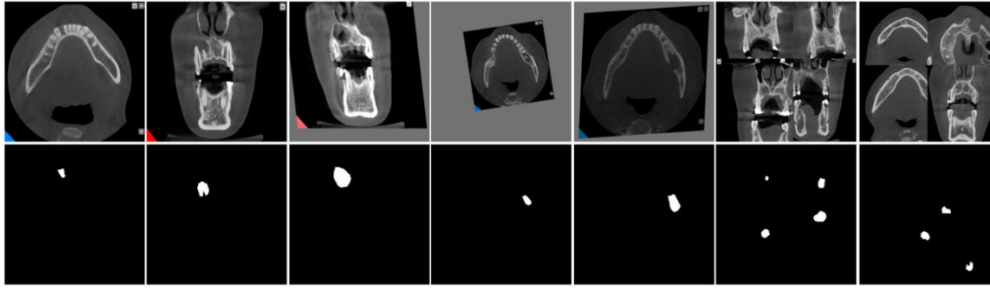


FIGURE 7. Some original and augmented sample images of jaw cyst (top) with their corresponding labels (bottom).

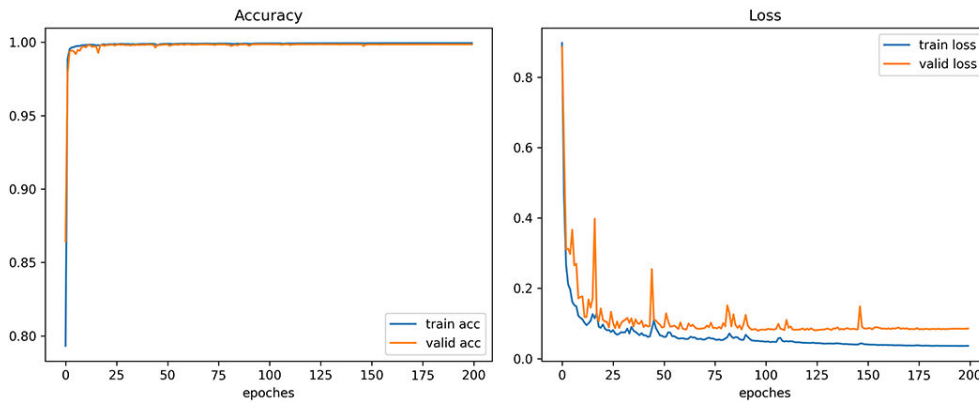


FIGURE 8. Accuracy and loss curve of the proposed MFI-Net on the original jaw cyst dataset.

TABLE 1. Optimizer selection experiment of the proposed MFI-Net.

Optimizer	Dice (%)	IoU (%)	Jaccard (%)
Adamax	90.55	91.40	82.97
AdamW	92.80	93.25	86.61
NAdam	92.71	93.18	86.47
RMSPro	92.35	92.90	85.92
Rprop	92.18	92.72	85.58
SGD	91.45	92.13	84.41
Adam	93.06	93.47	87.06

augmented dataset, we enhanced the training dataset and validation dataset to obtain 2765 for training, 920 for validation, and 917 for testing. Owing to limited space, Figure 7 only provides some original and augmented sample images of jaw cyst.

B. MFI-Net TRAINING AND VERIFICATION

The proposed network is implemented on windows 64-bit system using the robust capabilities of PyTorch 1.8.0 library and an NVIDIA Quadro RTX 6000 graphic card, boasting an expansive 24 GB memory capacity. Subsequently, to optimize the model weights more reasonably, we adopt the Adam as optimizer. In addition, some additional hyper-parameters are set: the batch size is 16, the number of iterations is 200, and the initial learning rate is 0.001. To mitigate the detrimental effects of model over-fitting, we employ a sophisticated strategy known as early stopping. When the performance on the

validation set is no longer increasing, we will stop the training process to prevent the model from going deep into the over-fitting region. Figure 8 is the accuracy and loss curve of the proposed MFI-Net on the original jaw cyst dataset, where the blue represents the training curve and the orange represents the validation curve. It can be observed from the graphical data that the MFI-Net achieves superior results in both the performance metrics and the convergence rates throughout the training and validation phases, without any over-fitting or underfitting.

C. EVALUATION METRICS

In this paper, three vital indicators, including Dice [36], [37], Intersection over union (IoU) [38], [39] and Jaccard [40], [41], are employed to assess the performance of various models, the details of which are calculated as:

$$Dice = \frac{2TP}{2TP + FN + FP} \quad (4)$$

$$IoU = \frac{TP}{TP + FN + FP} \quad (5)$$

$$Jaccard = \frac{TP}{TP + FN + FP} \quad (6)$$

where TP and FP refer to true positive and false positive, TN and FN stand for true negative and false negative.

TABLE 2. Comparison experiment on the original jaw cyst dataset.

Method	Dice (%)	IoU (%)	Jaccard (%)	Params(M)	FPS
U-Net [15]	90.62	91.49	83.13	1.94	252.18
DABNet [42]	91.55	92.20	84.54	0.75	108.20
EDANet [43]	90.59	91.53	83.20	0.68	15.32
BEMNet [44]	91.20	92.05	84.24	13.41	73.97
META-Unet [45]	90.73	91.63	83.42	21.70	88.07
AMFU-net [46]	91.85	92.45	85.04	0.47	62.72
OAU-net [47]	91.69	92.38	84.91	26.14	82.51
LFU-Net [48]	90.66	91.47	83.10	0.19	107.03
GCHA-Net [49]	92.37	92.91	85.94	4.41	165.92
MFI-Net	93.06	93.47	87.06	0.99	106.21

D. OPTIMIZER SELECTION

The selection of optimizer is critical as it influences how quickly and accurately the model can converge to a solution, especially in tasks requiring precise optimization. As shown in Table 1, it is the experimental results for the different optimizers (including Adamax, AdamW, NAdam, RMSPro, Rprop, SGD and Adam) of the proposed MFI-Net method on the original jaw cyst dataset. The results showed that Adamax has the lowest performance among the tested optimizers with 90.55% Dice, 91.40% IoU, and 82.97% Jaccard. The main reason is that the lack of adaptive learning rate limits the effectiveness of the optimizer in complex tasks. In terms of performance, the NAdam is very close to the AdamW, with both slightly underperforming the Adam. But Adam performed best overall, with Dice scores of 93.06%, IoU scores of 93.47% and Jaccard scores of 87.06%. Its adaptive learning rate mechanism, efficiency in handling sparse gradients, and ease of use make it the optimal choice for the MFI-Net model in our framework.

E. COMPARISON WITH OTHER METHODS

1) EXPERIMENTAL ON ORIGINAL DATASET

To underscore the superiority of our method, we utilize the original jaw cyst dataset to make meticulous comparisons with a number of architectures such as U-Net, DABNet, EDANet, BEMNet, META-Unet, AMFU-net, OAU-net, LFU-Net, and GCHA-Net. These models represent the vanguard of segmentation technology and have been carefully evaluated, with their performance intricately detailed in Tables 2. It's noteworthy that U-Net, EDANet, META-Unet, and LFU-Net exhibited relatively poor performance, indicating they were unable to process the complexity of jaw cyst images. Conversely, GCHA-Net's ingenious integration of both global and local attention mechanisms that allow it to recognize overall structural features and complex details. Compared with these nine networks with superior performance, MFI-Net emerges as a standout performer, boasting Dice, IoU, and Jaccard scores of 93.06%, 93.47%, and 87.06%, respectively. These metrics represent a significant improvement over U-Net, with margins of 2.44%, 1.98%, and 3.93%. The experimental results show that the proposed SE-Res2Conv, MPB, PAM, and MFIB are effective, they can

enhance the ability of the network to obtain feature information and improve the segmentation accuracy.

To provide an intuitive comparison of segmentation performance, the visualization results of our approach and various typical methods are shown in Figure 9. As can be seen from the figure, U-Net exhibits notable inaccuracies in segmenting small jaw cysts, leading to discontinuous segmentation (the third row of Figure 9). Inspiration from dilated convolution and dense connectivity, EDANet and LFU-Net achieve performance comparable to U-Net (the fifth and tenth rows of Figure 9). However, it is still fall short due to limitations in receptive field and multi-scale feature extraction. Addressing these challenges, DABNet regarded as a variant of U-Net, is developed by depthwise asymmetric bottleneck, to enhance the receptive field and exploit contextual information effectively, as demonstrated in the fourth row of Figure 9. More recently, BEMNet has further guided network segmentation by introducing a new boundary enhancement encoder-decoder into U-Net (the sixth row of Figure 9). To further refine jaw cyst segmentation, AMFU-net and OAU-net integrate residual attention blocks to expand the receptive field. However, see the eighth and ninth rows of Figure 9, there is discontinuity and false segmentation in the segmentation results of AMFU-net and OAU-net, which is caused by insufficient semantic and global context information during the up-sampling process. In eleventh row of Figure 9, GCHA-Net achieved the second-best results by leveraging global and local attention mechanisms. In contrast, our approach is able to explicitly extend the receptive field and utilize multi-scale feature maps, which are largely thanks to our proposed SE-Res2Conv, MPB, and PAM. This allows the comprehensive use of semantic information at different scales and facilitates the fusion of different scales through feature mapping weighting. Additionally, the multi-level feature integration block efficiently combines intricate low-level details with broader high-level semantic features, culminating in superior segmentation results compared to other methods, as shown in the last row of Figure 9.

2) EXPERIMENTAL ON AUGMENTED DATASET

Furthermore, we further validated the performance of our proposed MFI-Net method by conducting extensive

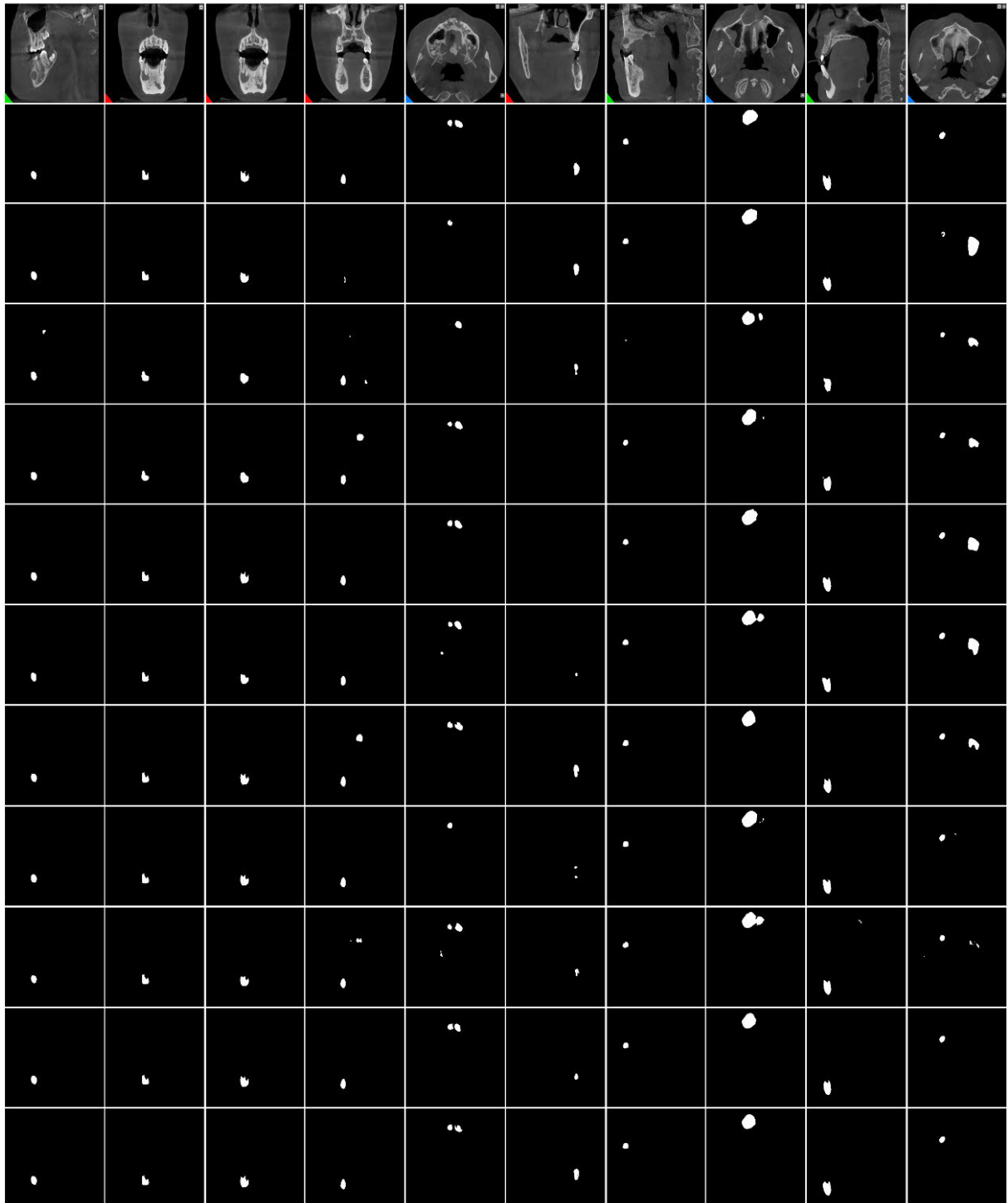


FIGURE 9. Visual segmentation results of different methods on the original jaw cyst dataset. The first and second rows: original images and their corresponding ground truth. The third to last rows are the results of U-Net, DABNet, EDANet, BEMNet, META-Unet, AMFU-net, OAU-net, LFU-Net, GCHA-Net and MFI-Net.

experiments on the augmented jaw cyst dataset. Qualitative visualizations are depicted in Figure 10 and segmentation outcomes are thoroughly presented in Table 3. It could be easily seen that the expansion of the dataset introduces unique challenges to image segmentation, leading

to instances of mis-segmentation and missing segmentation across all models to varying degrees. Consequently, there is a noticeable reduction in their respective Dice, IoU, and Jaccard values. Despite these challenges, our proposed method outperforms all competing approaches across all

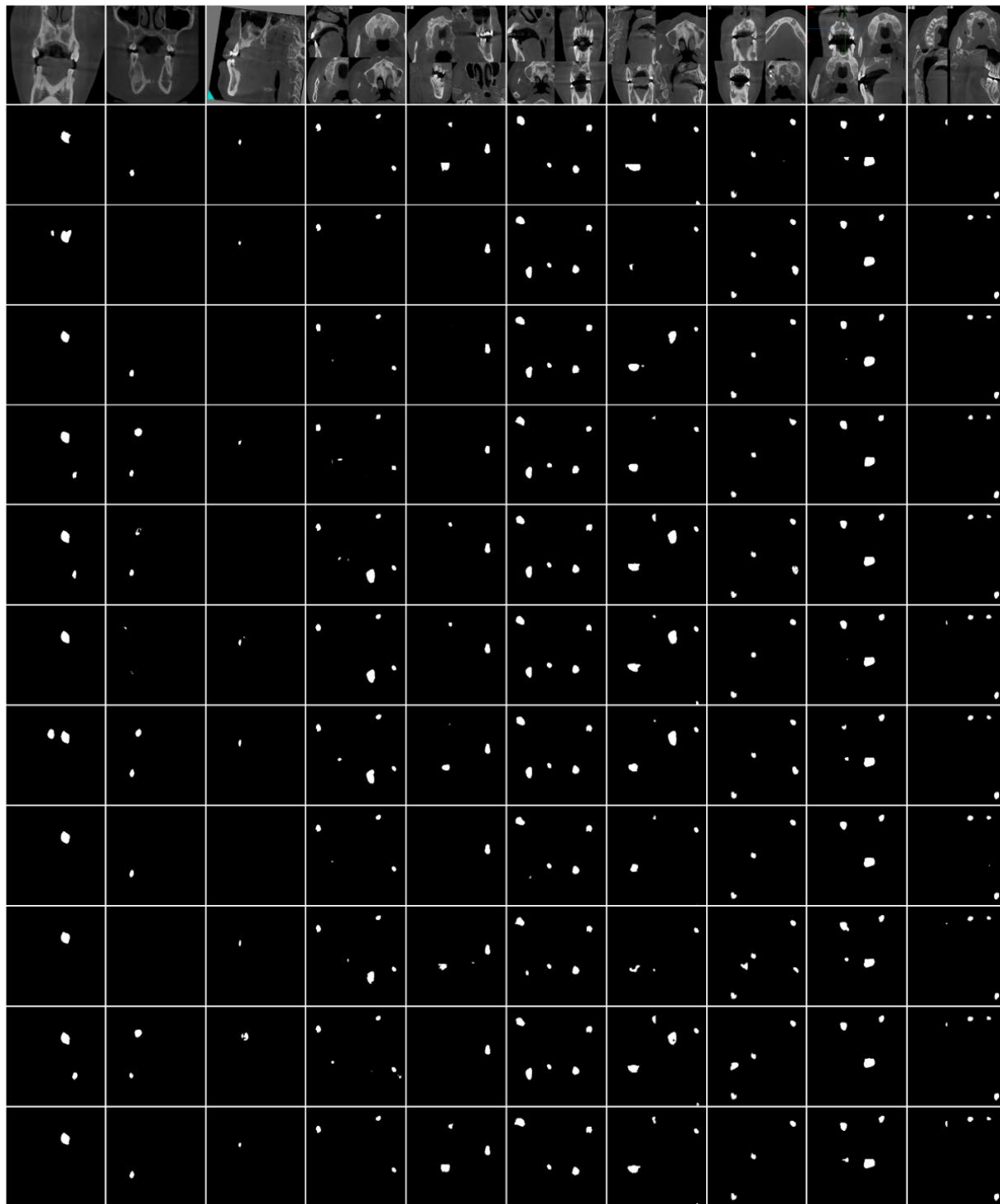


FIGURE 10. Visual segmentation results of different methods on the augmented jaw cyst dataset. The first and second rows: original images and their corresponding ground truth. The third to last rows are the results of U-Net, DABNet, EDANet, BEMNet, META-Unet, AMFU-net, OAU-net, LFU-Net, GCHA-Net and MFI-Net.

evaluation metrics, boasting an impressive Dice coefficient of 91.25%, IoU score of 91.94%, and Jaccard index of 84.06%. This superior performance underscores the robustness of the MFI-Net method and highlights its accuracy in

efficiently handling the complexity of augmented datasets. Thus, it demonstrates its superiority in accurately delineating the boundaries of jaw cysts amidst the augmented data variability.

TABLE 3. Comparison experiment on the augmented jaw cyst dataset.

Method	Dice (%)	IoU (%)	Jaccard (%)	Params(M)	FPS
U-Net [15]	89.58	90.58	81.37	1.94	254.54
DABNet [42]	90.50	91.29	82.79	0.75	115.14
EDANet [43]	89.26	90.29	80.81	0.68	14.58
BEMNet [44]	90.61	91.43	83.05	13.41	73.63
META-Unet [45]	90.35	91.19	82.58	21.70	92.24
AMFU-net [46]	90.12	90.97	82.16	0.47	61.01
OAU-net [47]	91.14	91.84	83.85	26.18	72.79
LFU-Net [48]	88.13	89.36	78.97	0.19	105.96
GCHA-Net [49]	91.00	91.71	83.61	4.41	162.96
MFI-Net	91.25	91.94	84.06	0.99	110.28

3) COMPLEXITY AND OPERATIONAL EFFICIENCY

In addition, we utilize two key metrics: parameters (Params) and frame per second (FPS) to evaluate the efficiency of the above methods. These metrics are key for evaluating the effectiveness and practical applicability of different neural network architectures, while also taking into account the complexity of the model and the speed of computation in the inference process. As listed in Table 2 and 3, As shown in Tables 2 and 3, our MFI-Net shows a significant balance between model complexity and inference speed. By using relatively few parameters, our model achieved high FPS rates of 106.21 and 110.28 on both the original jaw cyst dataset and the augmented dataset. Although slightly slower than models like U-Net, DABNet, and GCHA-Net, our MFI-Net offers significant performance improvements. Considering the substantial advances in segmentation accuracy and overall model robustness, it is considered acceptable.

F. ABLATION STUDIES

To delve more profoundly into the understanding of the SE-Res2Conv, MPB+PAM, and the multi-level feature integration block, we use ablation experiments to dissect the individual contributions of these modules. Initially, establish the original U-Net architecture as our baseline. Subsequently, we integrate the aforementioned components one by one. Finally, the Dice, IoU and Jaccard are employed for performance evaluation. Both quantitative and visual results on the original jaw cyst dataset are presented in Table 4 and Figure 11, providing a comprehensive overview of our findings.

1) EFFICACY OF SE-Res2Conv

Firstly, the proposed SE-Res2Conv is integrated into the Baseline, with a comprehensive visualization showcased in the third and fourth rows of Figure 11. This addition notably enhances the segmentation accuracy of the Baseline, as evidenced by its ability to capture a more extensive array of vessel pixels. From Table 4, we can confirm that after adding SE-Res2Conv to the Baseline (Baseline + SE-Res2Conv), all evaluation metrics have improved to a certain extent. Notably, when compared to the Baseline, the Dice, IoU, and Jaccard scores improve from 90.62%, 91.49%, 83.13 to 91.60%, 92.38%, and 84.89%, increasing by 0.98%, 0.89%,

and 1.76%, respectively. The main reason is attributed to the introduction of SE-Res2Conv as an encoder, which can better maintain model's capacity in extracting features across various scales effectively.

2) EFFICACY OF MPB+PAM

In the second stage, we incorporate the MPB and PAM into the context extractor module to replace the traditional convolutional layers. As shown in Table 4, there has been a significant improvement in segmentation performance. In addition, a more detailed evaluation of the segmentation results is provided through visual representations in the fifth row of Figure 11. It is worth noting that the Baseline model enhanced by MPB+PAM shows significant proficiency in segmenting large and small objects, especially in depicting cysts characterized by large-scale changes in low contrast areas. Overall, the combination of the Baseline model with the (MPB+PAM) configuration has been proven to be very effective in our framework.

3) EFFICACY OF MFIB

Third, we introduce the MFIB into Baseline architecture, denoted as (Baseline + MFIB), with the aim of evaluating its effectiveness. Since the Baseline network lies in its utilization of simple jump connections at each layer to depict local information, the semantic information cannot be fully explored. As illustrated in the sixth row of Figure 11 and a comprehensive analysis presented in Table 4, by seamlessly integrating semantic concepts and spatial details, our MFIB can enable networks to extract subtle differences from complex data with greater efficiency.

4) EFFICACY OF FUSION MODULE

Finally, to effectively convey contextual information, we design a fusion module (Baseline + SE-Res2Conv + (MPB+PAM) + MFIB) by combining SE-RES2CONV, MPB+PAM and MFIB. As seen from the last row of Figure 11, compared with the Baseline network, our approach is able to capture relatively complete topology and refined segmentation results. As shown in Table 4, our approach shows tremendous improvement in the Dice, IoU, and Jaccard scores, which is 2.46%, 1.98% and 3.93%, respectively, compared to the Baseline network. As can be seen from the

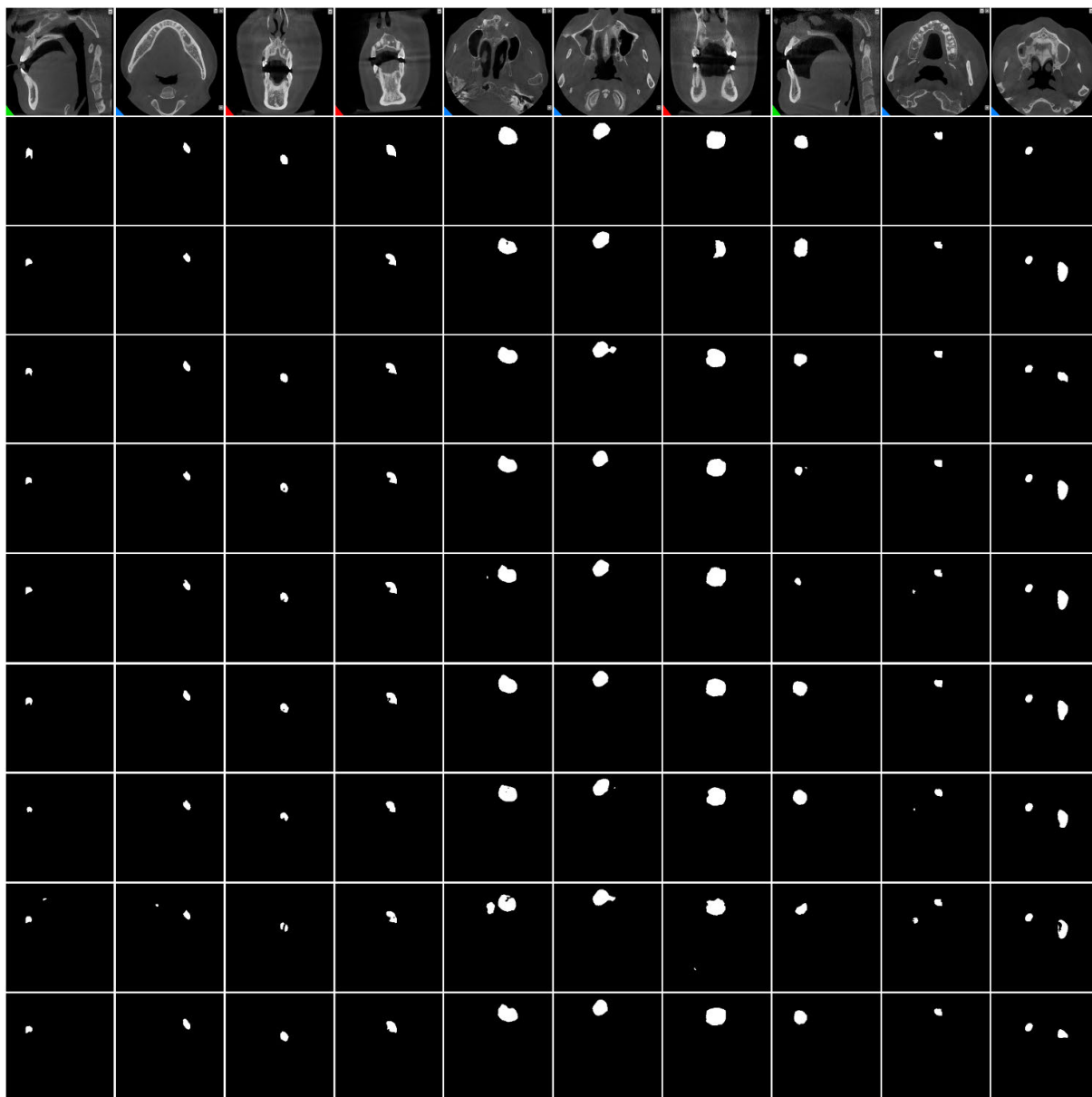


FIGURE 11. Visualization results of ablation studies. The first and second rows: original images and their corresponding ground truth. The third to last rows are the results of Baseline, Baseline + SE-Res2Conv, Baseline + (MPB+PAM), Baseline + MFIB, Baseline + SE-Res2Conv + (MPB+PAM), Baseline + SE-Res2Conv + MFIB, Baseline + (MPB+PAM) + MFIB, and Baseline + SE-Res2Conv + (MPB+PAM) + MFIB.

TABLE 4. Ablation experiment on the original jaw cyst dataset.

Model	Dice (%)	IoU (%)	Jaccard (%)
Baseline	90.62	91.49	83.13
Baseline + SE-Res2Conv	91.60	92.38	84.89
Baseline + (MPB+PAM)	92.06	92.67	85.46
Baseline + MFIB	91.68	92.36	84.86
Baseline + SE-Res2Conv + (MPB+PAM)	92.48	92.99	86.12
Baseline + SE-Res2Conv + MFIB	91.97	92.54	85.21
Baseline + (MPB+PAM) + MFIB	92.59	93.08	86.28
Baseline + SE-Res2Conv + (MPB+PAM) + MFIB	93.06	93.47	87.06

visual and statistical results, each component in our model is effective, and the best segmentation results can be obtained

by combining these components together. Therefore, the proposed approach is well suited for jaw cyst segmentation.

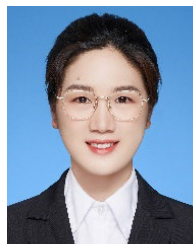
IV. CONCLUSION

To address the complex difficulties related to insufficient contextual information and the loss of details pertaining to jaw cyst, a new multi-level feature integration network is developed. Our framework consists of three primary components: encoder module, context extractor module and decoder module. In particular, by employing the SE-Res2Conv structure, the capacity to capture information across various scale receptive fields has been significantly improved. Furthermore, integrating MPB with PAM resulted in the generation of features with greater distinctiveness. Lastly, a multi-level feature integration block has been devised to efficiently integrate low-level detail features with high-level semantic features. The effectiveness of MFI-Net has been confirmed through evaluation on newly constructed original and augmented jaw cyst datasets, where it demonstrated superior performance over other contemporary state-of-the-art segmentation techniques.

REFERENCES

- [1] A. S. Ding, A. Lu, Z. Li, D. Galaiya, M. Ishii, J. H. Siewerdsen, R. H. Taylor, and F. X. Creighton, "Statistical shape model of the temporal bone using segmentation propagation," *Otol. Neurotol.*, vol. 43, no. 6, pp. e679–e687, 2022.
- [2] A. Afzali, F. Babapour Mofrad, and M. Pouladian, "2D statistical lung shape analysis using chest radiographs: Modelling and segmentation," *J. Digit. Imag.*, vol. 34, no. 3, pp. 523–540, 2021.
- [3] R. Srikanth and K. Bikshalu, "Chaotic multi verse improved Harris hawks optimization (CMV-IHHO) facilitated multiple level set model with an ideal energy active contour for an effective medical image segmentation," *Multimedia Tools Appl.*, vol. 81, no. 15, pp. 20963–20992, Jun. 2022.
- [4] Q. Cai, Y. Qian, S. Zhou, J. Li, Y.-H. Yang, F. Wu, and D. Zhang, "AVLSM: Adaptive variational level set model for image segmentation in the presence of severe intensity inhomogeneity and high noise," *IEEE Trans. Image Process.*, vol. 31, pp. 43–57, 2022.
- [5] P. Peng, D. Wu, F.-C. Han, L.-J. Huang, Z. Wei, J. Wang, Y. Jiang, and K. Xia, "Segmentation of breast molybdenum target image lesions based on semi-supervised fuzzy clustering," *J. Intell. Fuzzy Syst.*, vol. 44, no. 6, pp. 9475–9493, Jun. 2023.
- [6] N. A. Ali, A. El Abbassi, and O. Bouattane, "Performance evaluation of spatial fuzzy C-means clustering algorithm on GPU for image segmentation," *Multimedia Tools Appl.*, vol. 82, no. 5, pp. 6787–6805, Feb. 2023.
- [7] Z. Zhang, H. Wu, H. Zhao, Y. Shi, J. Wang, H. Bai, and B. Sun, "A novel deep learning model for medical image segmentation with convolutional neural network and transformer," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 15, no. 4, pp. 663–677, Dec. 2023.
- [8] X. Liu, L. Yang, J. Chen, S. Yu, and K. Li, "Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103165.
- [9] Y. Peng, L. Pan, P. Luan, H. Tu, and X. Li, "Curvilinear object segmentation in medical images based on ODoS filter and deep learning network," *Appl. Intell.*, vol. 53, no. 20, pp. 23470–23481, Oct. 2023.
- [10] H. Messaoudi, A. Belaid, D. Ben Salem, and P.-H. Conze, "Cross-dimensional transfer learning in medical image segmentation with deep learning," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102868.
- [11] T. S. Arulananth, P. G. Kuppusamy, R. K. Ayyasamy, S. M. Alhashmi, M. Mahalakshmi, K. Vasanth, and P. Chinnasamy, "Semantic segmentation of urban environments: Leveraging U-Net deep learning model for cityscape image analysis," *PLoS ONE*, vol. 19, no. 4, Apr. 2024, Art. no. e0300767.
- [12] T. S. Arulananth, S. W. Prakash, R. K. Ayyasamy, V. P. Kavitha, P. G. Kuppusamy, and P. Chinnasamy, "Classification of paediatric pneumonia using modified DenseNet-121 deep-learning model," *IEEE Access*, vol. 12, pp. 35716–35727, 2024.
- [13] L. Xu, K. Qiu, K. Li, G. Ying, X. Huang, and X. Zhu, "Automatic segmentation of ameloblastoma on ct images using deep learning with limited data," *BMC Oral Health*, vol. 24, no. 1, p. 55, Jan. 2024.
- [14] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
- [15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.
- [16] J. Qiao, X. Wang, J. Chen, and M. Liu, "MBUTransNet: Multi-branch U-shaped network fusion transformer architecture for medical image segmentation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 18, no. 10, pp. 1895–1902, Apr. 2023.
- [17] K. Fang, B. He, L. Liu, H. Hu, C. Fang, X. Huang, and F. Jia, "UMRFormer-Net: A three-dimensional U-shaped pancreas segmentation method based on a double-layer bridged transformer network," *Quant. Imag. Med. Surgery*, vol. 13, no. 3, pp. 1619–1630, Mar. 2023.
- [18] Q. Sun, M. Dai, Z. Lan, F. Cai, L. Wei, C. Yang, and R. Chen, "UCR-Net: U-shaped context residual network for medical image segmentation," *Comput. Biol. Med.*, vol. 151, Dec. 2022, Art. no. 106203.
- [19] Y. Arijji, K. Araki, M. Fukuda, M. Nozawa, C. Kuwada, Y. Kise, and E. Arijji, "Effects of the combined use of segmentation or detection models on the deep learning classification performance for cyst-like lesions of the jaws on panoramic radiographs: Preliminary research," *Oral Sci. Int.*, vol. 21, no. 2, pp. 198–206, May 2024.
- [20] Q. Tan, M. Ye, A. J. Ma, B. Yang, T. C. Yip, G. L. Wong, and P. C. Yuen, "Explainable uncertainty-aware convolutional recurrent neural network for irregular medical time series," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4665–4679, Oct. 2021.
- [21] P. Chen, S. Huang, and Q. Yue, "Skin lesion segmentation using recurrent attentional convolutional networks," *IEEE Access*, vol. 10, pp. 94007–94018, 2022.
- [22] B. Goyal, D. C. Lepcha, A. Dogra, and S.-H. Wang, "A weighted least squares optimisation strategy for medical image super resolution via multi-scale convolutional neural networks for healthcare applications," *Complex Intell. Syst.*, vol. 8, no. 4, pp. 3089–3104, Aug. 2022.
- [23] B. Dourthe, N. Shaikh, A. Pai S., S. Fels, S. H. M. Brown, D. R. Wilson, J. Street, and T. R. Oxland, "Automated segmentation of spinal muscles from upright open MRI using a multiscale pyramid 2D convolutional neural network," *Spine*, vol. 47, no. 16, pp. 1179–1186, 2022.
- [24] A. Selvaraj and E. Nithiyaraj, "CEDRNN: A convolutional encoder-decoder residual neural network for liver tumour segmentation," *Neural Process. Lett.*, vol. 55, no. 2, pp. 1605–1624, Apr. 2023.
- [25] B. C. Anil and P. Dayananda, "Automatic liver tumor segmentation based on multi-level deep convolutional networks and fractal residual network," *IETE J. Res.*, vol. 69, no. 4, pp. 1925–1933, May 2023.
- [26] T. Kanimozhi and F. J. Vijay, "An automated cervical cancer detection scheme using deeply supervised shuffle attention modified convolutional neural network model," *Automatika*, vol. 64, no. 3, pp. 518–528, Jul. 2023.
- [27] R. Rasti, A. Biglari, M. Rezapourian, Z. Yang, and S. Farsiu, "RetiFluidNet: A self-adaptive and multi-attention deep convolutional network for retinal OCT fluid segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1413–1423, May 2023.
- [28] F. Zhang, S. Yan, Y. Zhao, Y. Gao, Z. Li, and X. Lu, "Iterative convolutional encoder-decoder network with multi-scale context learning for liver segmentation," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2151186.
- [29] B. Li, Y. Cong, and H. Mo, "Ophthalmic OCT segmentation method based on RCNN-attention," *IEEE Access*, vol. 11, pp. 129601–129612, 2023.
- [30] M. Mubashar, H. Ali, C. Grönlund, and S. Azmat, "R2U++: A multiscale recurrent residual U-Net with dense skip connections for medical image segmentation," *Neural Comput. Appl.*, vol. 34, no. 20, pp. 17723–17739, Oct. 2022.
- [31] L. Xiong, C. Yi, Q. Xiong, and S. Jiang, "SEA-NET: Medical image segmentation network based on spiral squeeze-and-excitation and attention modules," *BMC Med. Imag.*, vol. 24, no. 1, p. 17, Jan. 2024.
- [32] H. Üzen, M. Turkoglu, M. Aslan, and D. Hanbay, "Depth-wise squeeze and excitation block-based efficient-unet model for surface defect detection," *Vis. Comput.*, vol. 39, no. 5, pp. 1745–1764, May 2023.
- [33] M. K. Dhar, T. Zhang, Y. Patel, S. Gopalakrishnan, and Z. Yu, "FUSegNet: A deep convolutional neural network for foot ulcer segmentation," *Biomed. Signal Process. Control*, vol. 92, Jun. 2024, Art. no. 106057.

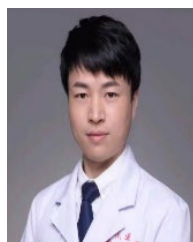
- [34] Z. Yu, L. Yu, W. Zheng, and S. Wang, "EIU-Net: Enhanced feature extraction and improved skip connections in U-Net for skin lesion segmentation," *Comput. Biol. Med.*, vol. 162, Aug. 2023, Art. no. 107081.
- [35] A. Sharma and P. K. Mishra, "DRI-UNet: Dense residual-inception UNet for nuclei identification in microscopy cell images," *Neural Comput. Appl.*, vol. 35, no. 26, pp. 19187–19220, Sep. 2023.
- [36] H. Tang, Y. Chen, T. Wang, Y. Zhou, L. Zhao, Q. Gao, M. Du, T. Tan, X. Zhang, and T. Tong, "HTC-Net: A hybrid CNN-transformer framework for medical image segmentation," *Biomed. Signal Process. Control*, vol. 88, Feb. 2024, Art. no. 105605.
- [37] R. Wu, P. Liang, X. Huang, L. Shi, Y. Gu, H. Zhu, and Q. Chang, "MHOrUNet: High-order spatial interaction UNet for skin lesion segmentation," *Biomed. Signal Process. Control*, vol. 88, Feb. 2024, Art. no. 105517.
- [38] D.-H.-N. Nham, M.-N. Trinh, V.-D. Nguyen, V.-T. Pham, and T.-T. Tran, "An EfficientNet-encoder U-Net joint residual refinement module with Tversky–Kahneman Baroni–Urbani–Buser loss for biomedical image segmentation," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104631.
- [39] Z.-U.-D. Muhammad, Z. Huang, N. Gu, and U. Muhammad, "DCANet: Deep context attention network for automatic polyp segmentation," *Vis. Comput.*, vol. 39, no. 11, pp. 5513–5525, Nov. 2023.
- [40] H. Abdel-Nabi, M. Z. Ali, and A. Awajan, "A multi-scale 3-stacked-layer coned U-Net framework for tumor segmentation in whole slide images," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105273.
- [41] K. Hu, Y. Zhu, T. Zhou, Y. Zhang, C. Cao, F. Xiao, and X. Gao, "DSC-Net: A novel interactive two-stream network by combining transformer and CNN for ultrasound image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.
- [42] G. Li, I. Yun, J. Kim, and J. Kim, "DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," 2019, *arXiv:1907.11357*.
- [43] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proc. ACM Multimedia Asia*, Dec. 2019, pp. 1–6.
- [44] W. Wu, Y. Gong, H. Hao, J. Zhang, P. Su, Q. Yan, Y. Ma, and Y. Zhao, "Choroidal layer segmentation in OCT images by a boundary enhancement network," *Frontiers Cell Develop. Biol.*, vol. 10, Nov. 2022, Art. no. 1060241.
- [45] H. Wu, Z. Zhao, and Z. Wang, "META-Unet: Multi-scale efficient transformer attention Unet for fast and high-accuracy polyp segmentation," *IEEE Trans. Autom. Sci. Eng.*, early access, 2023, doi: 10.1109/TASE.2023.3292373.
- [46] W. Y. Chung, I. H. Lee, and C. G. Park, "Lightweight infrared small target detection network using full-scale skip connection U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [47] H. Song, Y. Wang, S. Zeng, X. Guo, and Z. Li, "OAU-Net: Outlined attention U-Net for biomedical image segmentation," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104038.
- [48] Y. Deng, H. Wang, Y. Hou, S. Liang, and D. Zeng, "LFU-Net: A lightweight U-Net with full skip connections for medical image segmentation," *Current Med. Imag. Rev.*, vol. 19, no. 4, pp. 347–360, Apr. 2023.
- [49] H. Liu, Y. Fu, S. Zhang, J. Liu, Y. Wang, G. Wang, and J. Fang, "GCHA-Net: Global context and hybrid attention network for automatic liver segmentation," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106352.



HUIXIA ZHENG was born in Quzhou, China, in 1990. She received the master's degree in oral medicine from Zhejiang University. She is currently an Attending Physician with The Quzhou Hospital Affiliated of Wenzhou Medical University. Her research interests include diagnosis and imaging analysis of dental diseases.



XIAOLIANG JIANG received the M.S. and Ph.D. degrees in mechanical design from Southwest Jiaotong University, Chengdu, China. He is currently a Professor with Quzhou University. His research interests include machine vision and image recognition.



XU XU received the master's degree in oral medicine from Zhejiang University. He is currently the Chief Physician of The Quzhou Hospital Affiliated of Wenzhou Medical University. His research interests include diagnosis and treatment of dental diseases.



XIAOKANG DING received the M.S. and Ph.D. degrees in forest engineering from Beijing Forestry University. She is currently an Associate Professor with Quzhou University. Her research interests include machine vision and image recognition.

...