Mathematical Biosciences and Engineering

MBE, 21(3): 4036–4055.
DOI: 10.3934/mbe.2024178
Received: 17 November 2023
Revised: 14 January 2024
Accepted: 07 February 2024
Published: 23 February 2024

https://www.aimspress.com/journal/MBE

Research article

HIMS-Net: Horizontal-vertical interaction and multiple side-outputs network for cyst segmentation in jaw images

Xiaoliang Jiang[1,*,†], Huixia Zheng[2,*,†], Zhenfei Yuan[2], Kun Lan[1] and Yaoyang Wu[3]

[1] College of Mechanical Engineering, Quzhou University, Quzhou 324000, China
[2] Department of Stomatology, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou 324000, China
[3] Department of Computer and Information Science, University of Macau, Macau 999078, China

* **Correspondence:** Email: jxl_swjtu@163.com, qz_zhenghx@163.com.

† The authors contributed equally to this work.

Abstract: Jaw cysts are mainly caused by abnormal tooth development, chronic oral inflammation, or jaw damage, which may lead to facial swelling, deformity, tooth loss, and other symptoms. Due to the diversity and complexity of cyst images, deep-learning algorithms still face many difficulties and challenges. In response to these problems, we present a horizontal-vertical interaction and multiple side-outputs network for cyst segmentation in jaw images. First, the horizontal-vertical interaction mechanism facilitates complex communication paths in the vertical and horizontal dimensions, and it has the ability to capture a wide range of context dependencies. Second, the feature-fused unit is introduced to adjust the network's receptive field, which enhances the ability of acquiring multi-scale context information. Third, the multiple side-outputs strategy intelligently combines feature maps to generate more accurate and detailed change maps. Finally, experiments were carried out on the self-established jaw cyst dataset and compared with different specialist physicians to evaluate its clinical usability. The research results indicate that the Matthews correlation coefficient (Mcc), Dice, and Jaccard of HIMS-Net were 93.61, 93.66 and 88.10% respectively, which may contribute to rapid and accurate diagnosis in clinical practice.

Keywords: jaw cyst; image segmentation; horizontal-vertical interaction; feature-fused unit; multiple side-outputs

1. Introduction

A jaw cyst is a kind of non-neoplastic lesion, which is usually formed gradually by the formation of the epithelial tissue in the jawbone under certain conditions. During the pathological process, the cyst fluid continuously seeps out, causing a series of clinical symptoms such as facial swelling, tooth loosening, pathological fracture, and skin numbness. At present, the diagnosis of jaw diseases mainly depends on histopathological examination and the clinician's judgment of image results. The former is an invasive examination, while the latter, as a non-invasive examination, has the advantages of reducing patient discomfort, prevention of complications, faster recovery and less downtime, cost effectiveness, wider accessibility, repeatability, early detection, and disease surveillance, and it plays an important role in benign and malignant diagnosis, tumor staging, and boundary determination. In some cases, it is even possible to make a reliable diagnosis based on patient imaging. However, jaw cyst tissue often changes size and shape over time. Also, the tissue boundaries may not be well-defined and could be overlapping, which can be harsh and subjective for physicians. Some of the original images and their corresponding labels are shown in Figure 1. Therefore, it is necessary to find an efficient and accurate diagnosis method.

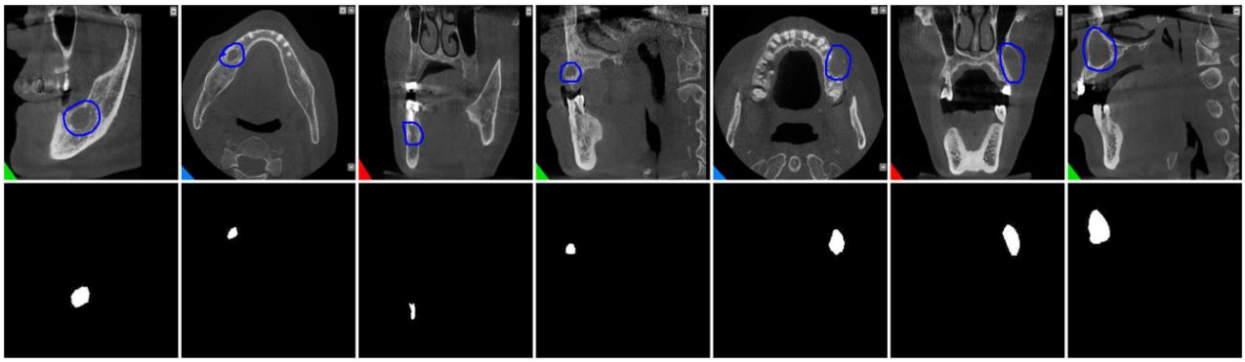

Figure 1. Some original sample images of jaw cysts and their corresponding labels. The original images (top) with their corresponding labels (bottom).

With the development of machine vision and imaging technology, deep-learning [1–5] plays an increasingly important role in the field of intelligent diagnosis, and a large number of algorithms have been applied to medical image segmentation. As the most well-known variant of full convolutional networks, U-Net [6] has a classic encoder-decoder symmetric structure. The encoder is used to identify and capture the contextual semantic information, while the decoder converts the high-level semantic information and the detailed features passed by the jump connection into the semantic labels required by the training. As a result, it greatly improves the performance of deep-learning. Since then, many improved U-Net architectures have been proposed, mainly based on residual mechanism [7,8], attention module [9,10], multi-scale features [11,12], and dilated convolution [13,14]. Among them, Vidal et al. [15] proposed a dynamic contrast enhancement framework based on U-Net to effectively solve the problem of class imbalance and confusing organs by using different input combinations and introducing residual basic blocks. Sun et al. [16] incorporated adaptive scaling block and feature refinement block into U-Net framework, which can better capture multi-scale features and channel dependencies, so the network obtained better

performance evaluation on DDSM-BCRP and INbreast databases. In addition, many algorithms have been proposed for the segmentation of jaw cyst images. Among them, Abdolali et al. [17] presented a segmentation approach based on asymmetry analysis that is versatile and applicable to various types of jaw cysts. It comprises three main steps: preprocessing and symmetry detection, image partition and correction, and intensity analysis with constraint enforcement. Alsmadi et al. [18] proposed a hybrid approach combining fuzzy c-means and neutrosophic techniques for segmenting jaw cysts in X-ray images, which may be helpful for early diagnosis of jaw lesions. Considering that transfer learning relies on the number of damaged samples and lacks reliability, He et al. [19] proposed a location-constrained dual network for the diagnosis of jaw cysts. Through self-supervision and pre-training of the feature extractor and the introduction of auxiliary segmentation branches to extract different features, the problem of data scarcity and reliability is effectively solved. Utilizing panoramic dental images, Sivasundaram et al. [20] proposed an improved LeNet for the classification of oral cyst images. Despite significant advances in existing methods, their fundamental reliance on convolutional computation inherently limits the ability to grasp global and remote features. Veena et al. [21] developed a geodesic active contour model for generating panoramic dental X-ray images, which plays a vital role in extracting relevant features that can greatly help clinicians for further analysis. Additionally, the existing approaches address inherent problems relating to a large number of training parameters. The complexity of these models often demands substantial computational resources, hindering their practical scalability. As these networks deepen, a phenomenon known as the vanishing gradient problem can occur, leading to deteriorating segmentation performance. These issues accentuate the need for more robust architectures that can mitigate these challenges and effectively learn both local details and broader, global context, ultimately enhancing the precision and efficiency of cyst segmentation in medical images.

In response to the above problems, an improved convolutional neural network framework is established to better segment cyst from jaw images. In our approach, there are three high-level blocks: horizontal-vertical interaction mechanism, feature-fused unit and multiple side-outputs strategy. The fusion of the above blocks has an efficient segmentation performance, which can achieve 93.61% Mcc, 93.66% Dice and 88.10% Jaccard. This paper has three specific contributions:

1) The horizontal-vertical interaction mechanism is adopted to make the network have stronger feature reuse capability without increasing parameters.

2) We propose a feature-fusion unit that utilizes extended convolution and standard convolution to obtain receptive fields with different sizes, so that the network can have richer context information.

3) The multi-side outputs strategy is used to fuse the feature information of different semantic levels. Moreover, the weighted loss function of binary cross entropy and Dice is utilized to improve the segmentation accuracy and reduce the influence of sample imbalance.

2. Materials and methods

2.1. Overview of HIMS-Net

Currently, there exists a large number of studies demonstrating the effectiveness of encoder-decoder architecture in medical image segmentation. Therefore, our HIMS-Net represents a noteworthy modification of the UNet++ [22,23] encoder-decoder structure, which is mainly composed of input layer, down-sampling layer, skip connect, feature-fusion unit layer, up-sampling

layer, and output layer, as shown in Figure 2. Each of these components plays a crucial role in the network's functioning. Specifically, the down-sampling layer serves to further compress the feature map through operations like maximum pooling or average pooling. Conversely, the up-sampling layer works to enlarge the feature map, restoring it to a higher resolution, which is vital for preserving intricate details in the output. Unlike the original UNet++, HIMS-Net eliminates dense connections at each stage and introduces the horizontal-vertical interaction mechanism that enables seamless information exchange between adjacent layers. This innovation enhances the network's ability to capture long-range dependencies within the data, thereby improving its performance. Then, the feature-fusion unit is used to replace traditional convolution kernel to obtain receptive fields with different sizes. This adaptation allows the network to better capture features at different scales and complexities within the input data. Moreover, there are multiple side-outputs at the top layer of the network, followed by a deep supervision layer. That is, each output branch is followed by a 1 × 1 convolutional layer. Finally, the loss function used for training HIMS-Net combines binary cross-entropy and Dice scores, which are weighted to facilitate lesion segmentation using the Softmax function. This combination of loss functions ensures that the network is trained to accurately segment lesions in the input data. For a more comprehensive understanding of each module and their specific functionalities, see the following subsections, where we will explain the HIMS-Net architecture in more depth.

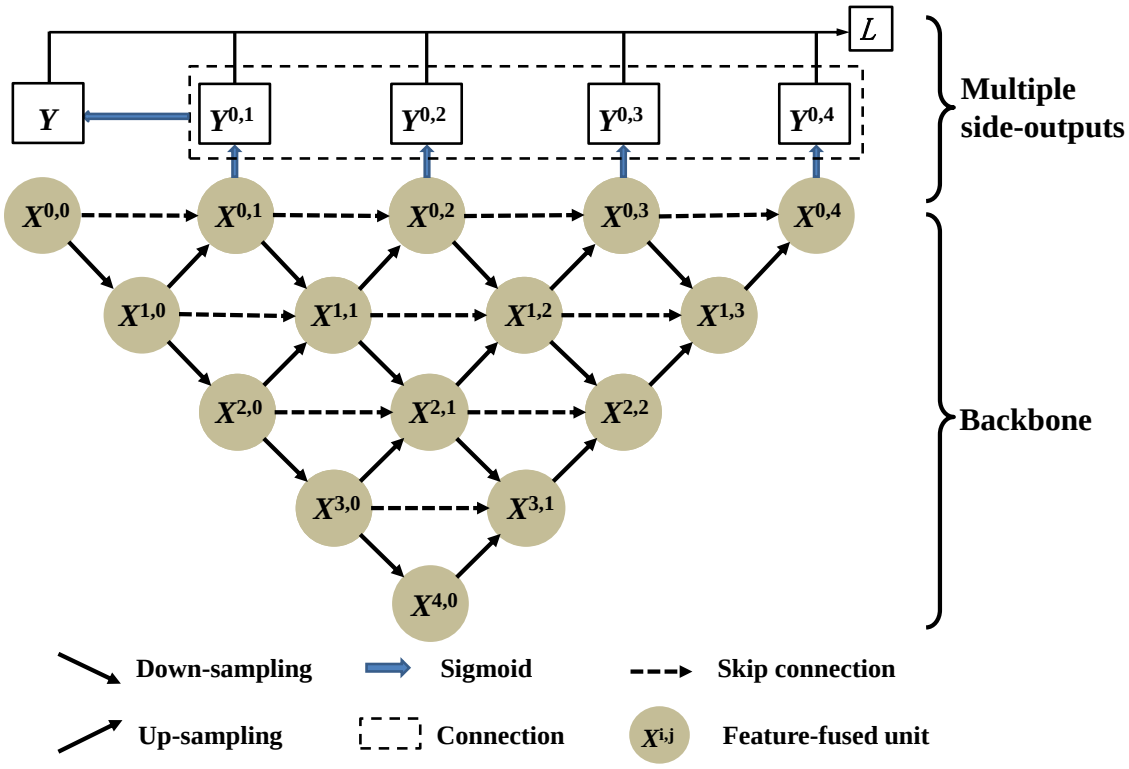

Figure 2. Architecture of HIMS-Net.

2.2. Horizontal-vertical interaction mechanism

In traditional architectures, communication is predominantly restricted to the forward and

backward propagation of data, leading to certain limitations in capturing long-range dependencies and holistic contextual information. The horizontal-vertical interaction mechanism embedded within the HIMS-Net architecture significantly influences information exchange across various network layers, playing a crucial role in preserving spatial details, enhancing feature reuse, and enabling the exchange of high-level semantic information and fine-grained details. This mechanism operates through both vertical and horizontal interactions, fostering the seamless flow of information. Vertically, it establishes connections between down-sampled and up-sampled layers, ensuring the propagation of information between lower-resolution compressed feature maps and higher-resolution detailed feature maps. Horizontally, within the same layer, lateral connections facilitate the exchange of long-range dependencies and lateral information among adjacent layers, allowing for comprehensive integration of global context and local details. Different from HRNet's [24] comprehensive high-to-low and low-to-high full connection, our approach specifically involves merging feature maps solely from adjacent stages. This interaction strategy enables the network to capture and process complex spatial structures in jaw images, contributing significantly to the accuracy of cyst segmentation without the unwarranted burden of augmenting network parameters and computational burden. Therefore, the horizontal-vertical interaction mechanism can be summarized into three ways, defined as follows:

$$x^{i,j} = \begin{cases} F(D(x^{i-1,j})), & j = 0 \\ F([x^{i,j-1}, U(x^{i+1,j-1})]), & i = 0 \\ F([D(x^{i-1,j}), x^{i,j-1}, U(x^{i+1,j-1})]), & other \end{cases} \tag{1}$$

where D and U are down-sampling and up-sampling operations, F is the feature-fusion unit, [ ] represents the concatenation layer, and $x^{i,j}$ denotes the output of node $X^{i,j}$. Specifically, node $j = 0$ only accepts input from the previous layer; node $i = 0$ receives two inputs from different layers; and for other nodes, the feature maps are three inputs from the previous layer, the same layer, and the next layer. In summary, the horizontal-vertical interaction mechanism within HIMS-Net represents a breakthrough approach that fosters intricate communication pathways across both vertical and horizontal dimensions of the network. By facilitating information exchange not only between adjacent layers but also among layers at the same resolution, it empowers the network to capture extensive contextual dependencies, ultimately enhancing the precision and accuracy of jaw cyst segmentation within medical imaging tasks.

2.3. Feature-fusion unit

In deep-learning methods (such as object detection and image segmentation), images often exhibit diverse structures and varying scales of relevant features. Specifically, low-level features have higher resolution and contain more position and detail information but have lower semantics and more noise due to fewer convolution operations. High-level features have stronger semantic information but the resolution is very low and the perception of detail is poor. Traditional convolutional neural networks face the challenge of adequately leveraging these multi-scale features to achieve accurate segmentation. To obtain richer context information, a feature-fusion unit is introduced to obtain receptive fields with different sizes, as shown in Figure 3. This unit operates as an advanced feature integration mechanism, aimed at enriching the network's capacity to capture

multi-scale contextual information essential for accurate segmentation. First, for a given feature map f_{in}, it branches out the feature map into two streams, utilizing 3 × 3 kernel filters with varying expansion sizes. These streams capture features at different receptive field scales, one focusing on a smaller field and the other on a larger field. Second, the results from the two branches were fused by summing the elements. After that, global averaging pooling (GAP), Dense, ReLU, Sigmoid, and Lambda were used. This sequence of operations aims to accentuate the inherent features captured from varying scales while minimizing noise and enhancing the network's ability to extract rich context information. Then, the above results were weighted with the feature graphs f_1 and f_2, and the output of the feature-fusion unit was calculated. In essence, the feature-fusion unit enriches the network's ability to capture contextual information across multiple scales, crucial for achieving accurate segmentation in jaw cyst images.

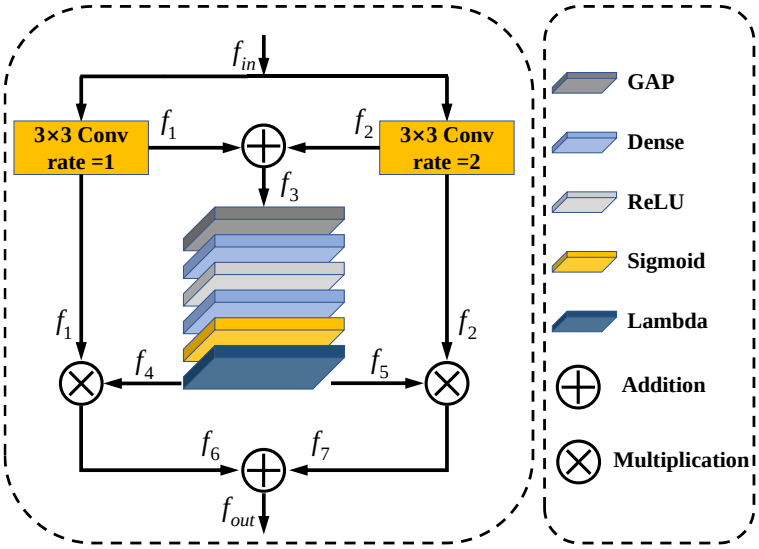

Figure 3. Structure of feature-fusion unit.

2.4. Multi-side outputs

For the HIMS-Net structure, the network depth continuously increases from left to right, and the output feature maps corresponding to each input image also becomes more and more refined. At the rightmost end of the network, the region segmentation of the target is best due to the increase of the convolutional layer. When making lesion prediction, the deepest network output is usually selected as the final result. However, sometimes the relatively shallow output can contain useful information, and even if the error area appears in the deepest output prediction result, other shallow output results may get the correct prediction result. Therefore, combining the useful information in shallow feature and deep feature is conducive to improving the prediction and segmentation accuracy of the whole network.

As shown in Figure 1, the feature maps generated by the four convolution units $\{X^{0,1}, X^{0,2}, X^{0,3}, X^{0,4}\}$ are obtained by a 1 × 1 convolution and Sigmoid function, and the results of the corresponding four side outputs are $\{Y^{0,1}, Y^{0,2}, Y^{0,3}, Y^{0,4}\}$. The new convolution unit Y generated by the four output results can be expressed as:

$$Y = Y^{0,1} \oplus Y^{0,2} \oplus Y^{0,3} \oplus Y^{0,4}, \tag{2}$$

where $\oplus$ represents the concatenation operation. Through the multi-side outputs fusion operation, the feature information from the four output layers will be fused into the final output Y, and the loss will be calculated according to the output result $\{Y^{0,1}, Y^{0,2}, Y^{0,3}, Y^{0,4}, Y\}$. By adjusting the network weight parameters through back-propagation, the output results of different depth layers can be optimized to different degrees, so that finer detection and identification details can be captured.

2.5. Loss function

The development and implementation of an appropriate loss function play a pivotal role in the training process of deep-learning networks, particularly in tasks like image segmentation. The loss function used in the HIMS-Net is a strategic combination of the binary cross-entropy loss and the Dice loss, amalgamated to maximize the network's accuracy and robustness in segmenting jaw cysts. Among them, binary cross entropy loss [25–28] is to predict the category of each pixel and then average all pixels. In essence, it is still equal learning for each pixel of the image, which reduces the feature extraction ability of non-mainstream categories if there is an imbalance in multiple categories of the image. On the other hand, Dice loss [29–32] considers all pixels in a category as a whole and calculates the proportion of the intersection in the whole, so it will not be affected by a large number of mainstream pixels and can extract better effects. Therefore, the HIMS-Net strategically combines the binary cross-entropy and Dice losses, emphasizing the advantages of both methods while compensating for their individual limitations. The weight of the binary cross entropy loss and the Dice loss is defined as:

$$L_{side}^{j} = L_{bce}^{j} + \mu L_{dice}^{j}, \tag{3}$$

where $j \in \{1, 2, 3, 4, 5\}$, L_{bce}^{j} represents the binary cross entropy loss, L_{dice}^{j} shows the dice loss, $\mu = 0.5$ is the weight coefficient of the two losses, and L_{side}^{j} represents the loss value from the jth output. Therefore, the total loss function L can be written as:

$$L = \sum_{i=1}^{5} \omega_i L_{side}^{j}, \tag{4}$$

where ω_i is the weight coefficient of each side-output layer.

3. Experiments and results

To verify the effectiveness of HIMS-Net for cyst segmentation in jaw images, thirteen common algorithms were used for comparison, including U-Net [6], UNet++ [22,23], UNet 3+ [33], FSP2-Net [34], SAR-U-Net [35], FCSNet [36], FF-UNet [37], DUDA-Net [38], BCDU-Net [39], UNet++-MSOF [40], META-Unet [41], MSU-Net [42], and CE-Net [43]. The image data involved in the experiment, sourced from the records of Quzhou People's Hospital, is a critical foundation for our exploration into jaw cyst segmentation within medical images. The dataset itself comprises a total

of 1535 images, of which 306 were used for testing, 922 for training, and 307 for verification. In order to facilitate rigorous experimentation and ensure the reliability of our findings, we imposed a strict standardization of image sizes across the entire dataset, each bearing the dimensions of 256 × 256 pixels. At the same time, all models were implemented based on Keras library and run on an NVIDIA RTX6000 graphics card with 24 GB of memory. In the training process, the batch-size was set to 16, initial learning rate was 0.001, the number of iterations was 200, and Adam optimization strategy [44] was adopted to achieve gradient descent. When the model has not improved beyond 50 epoch validation set metrics, iterative updates are stopped.

3.1. Evaluation metrics

Segmentation of a jaw cyst is essentially a pixel-level binary classification problem. If a pixel belongs to the cyst class, it corresponds to the label 1. If a pixel is a background class, its corresponding label is 0. The closer the predicted value is to 1, the higher the probability of a cyst and the lower the probability of a background. Therefore, in order to do quantitative analysis of these methods, Mcc [45], Dice [46], and Jaccard [47] are utilized, which can be given as:

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}, \tag{5}$$

$$Dice = \frac{2TP}{2TP + FN + FP}, \tag{6}$$

$$Jaccard = \frac{TP}{TP + FN + FP}, \tag{7}$$

where TP indicates that the judgment is positive and the fact is positive, TN means the judgment is negative and the fact is negative, FP indicates that the judgment is positive and the fact is negative, and FN means the judgment is negative and the fact is positive.

3.2. Parameter setting

3.2.1. Effect of optimizer

To determine the optimal optimizer to improve the performance of the proposed model, the power of the model was evaluated using a series of optimizers, including Adagrad, RMSProp, SGD, Adamax, and Adam, as shown in Table 1. Upon meticulous examination of the segmentation outcomes presented in the table, it becomes evident that the SGD optimizer displays the least favorable performance among the evaluated optimizers with Mcc, Dice and Jaccard of 90.51, 90.54 and 82.88%. Conversely, the Adam optimizer emerges as the standout performer, demonstrating exceptional segmentation performance with notably higher metrics of 93.50% for Mcc, 93.51% for Dice and 87.86% for Jaccard. The Adam optimizer's superiority in the context of jaw cyst segmentation with HIMS-Net can be attributed to its adaptive learning rates, momentum-based updates, and superior performance in achieving high accuracy. In contrast, Adagrad, RMSProp, SGD, and Adamax may face the challenge of navigating effectively in complex, high, and non-convex

optimized landscapes. Therefore, the robustness and efficiency in complex optimization landscapes are key factors that position Adam as the preferred optimizer for the specific task of jaw cyst segmentation in this study. This decision was informed by the clear advantages and superior performance displayed by Adam, indicating its robustness and efficiency in optimizing the proposed model for precise and effective segmentation tasks.

Table 1. HIMS-Net of different optimization algorithms.

Optimizer	Mcc	Dice	Jaccard
Adagrad	0.9210	0.9214	0.8552
RMSprop	0.9280	0.9281	0.8669
SGD	0.9051	0.9054	0.8288
Adamax	0.9164	0.9154	0.8479
Adam	**0.9350**	**0.9351**	**0.8786**

3.2.2. Effect of parameter ω_i

The loss function holds a pivotal role in the final outcomes of jaw cyst segmentation. In particular, the coefficient ω_i, which meticulously governs the weighting of loss term associated with each side-output layer. To gain a comprehensive understanding of the parameter's sensitivity and its impact on the segmentation outcomes, we conducted an extensive series of experiments with different values for ω_i. After analyzing the segmentation results, we selected the optimal combination $\{\omega_1 = 0.5, \omega_2 = 0.5, \omega_3 = 0.75, \omega_4 = 0.5, \omega_5 = 1.0\}$ according to the literature [32]. The experiment data details are shown in Tables 2.

Table 2. Mcc, Dice, and Jaccard of HIMS-Net with different values for ω_i.

ω_1	ω_2	ω_3	ω_4	ω_5	Mcc	Dice	Jaccard
0.05	0.05	0.05	0.05	0.8	0.9299	0.9300	0.8702
0.1	0.1	0.1	0.1	0.6	0.9288	0.9285	0.8689
0.15	0.15	0.15	0.15	0.4	0.9291	0.9293	0.8691
0.2	0.2	0.2	0.2	0.2	0.9289	0.9290	0.8690
0.5	0.5	0.5	0.5	0.5	0.9268	0.9265	0.8655
1.0	1.0	1.0	1.0	1.0	0.9316	0.9319	0.8732
0.5	**0.5**	**0.75**	**0.5**	**1.0**	**0.9350**	**0.9351**	**0.8786**

3.3. Results of HIMS-Net

The HIMS-Net network was trained on the jaw cyst data set, and its performance was tested under the test set. Figure 4 shows the loss function curve and accuracy curve of HIMS-Net during training and verification. It can be observed that in the initial training stage, with the increase of the number of iterations, the loss decreases rapidly and gradually becomes stable. This convergence, a key milestone in training, reflects the model's ability to adapt to the dataset, learning the underlying patterns and features essential for jaw cyst identification. In addition, the proposed network has a fast convergence rate, and the difference between the accuracy value of the model in the training set and

the accuracy value in the verification set is small, which indicates that the network has a good generalization performance. Thus, the stabilization of the loss function and accuracy metrics during the training process indicates a well-balanced learning trajectory, mitigating issues of over-fitting or under-fitting that could potentially compromise the model's performance on unseen data.

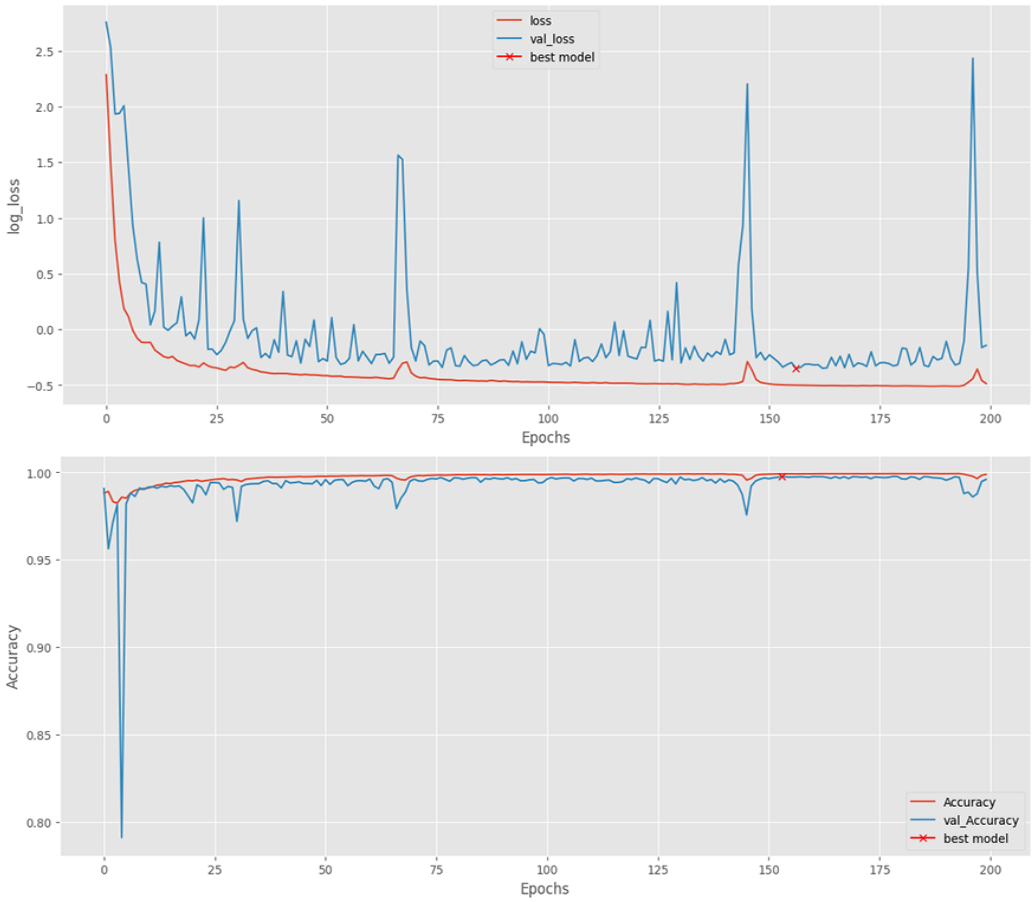

Figure 4. Each index of HIMS-Net during training and verification. Loss function (top) and accuracy curve (bottom).

3.4. Ablation experiments

Table 3. Ablation experiments of different structures.

Structure	Mcc	Dice	Jaccard
baseline	0.9064	0.9061	0.8307
Baseline + multiple side-outputs	0.9215	0.9213	0.8562
Baseline + feature-fusion unit	0.9194	0.9199	0.8521
Baseline + multiple side-outputs + feature-fusion	**0.9350**	**0.9351**	**0.8786**

In order to verify the validity of feature-fusion unit and multiple side-outputs, we used horizontal-vertical interaction network as the baseline model and then added modules to verify the outputs step by step. As shown in Table 3, by incorporating the horizontal-vertical interaction mechanism, the baseline method has significantly improved performance when compared to U-Net. From the third and fourth rows, it can be found that by adding multi-side outputs or feature-fusion

unit alone, evaluation metric of the model has been greatly increased. However, the most noteworthy enhancement was observed when HIMS-Net was structured with a combination of cascaded multiple side-outputs+feature-fusion design, showcasing improved performance across all evaluation indices. Moreover, a detailed comparison in Figure 5 distinctly illustrates the method's capacity to effectively mitigate false segmentations and exhibit superior connectivity. This visual representation emphasizes the method's exceptional ability to reduce inaccuracies in segmentation while ensuring better overall connectivity in the generated output. This observation strongly suggests that the incorporation of these additional modules significantly augments the method's capability to focus on crucial features, thereby contributing to heightened robustness in the segmentation process.

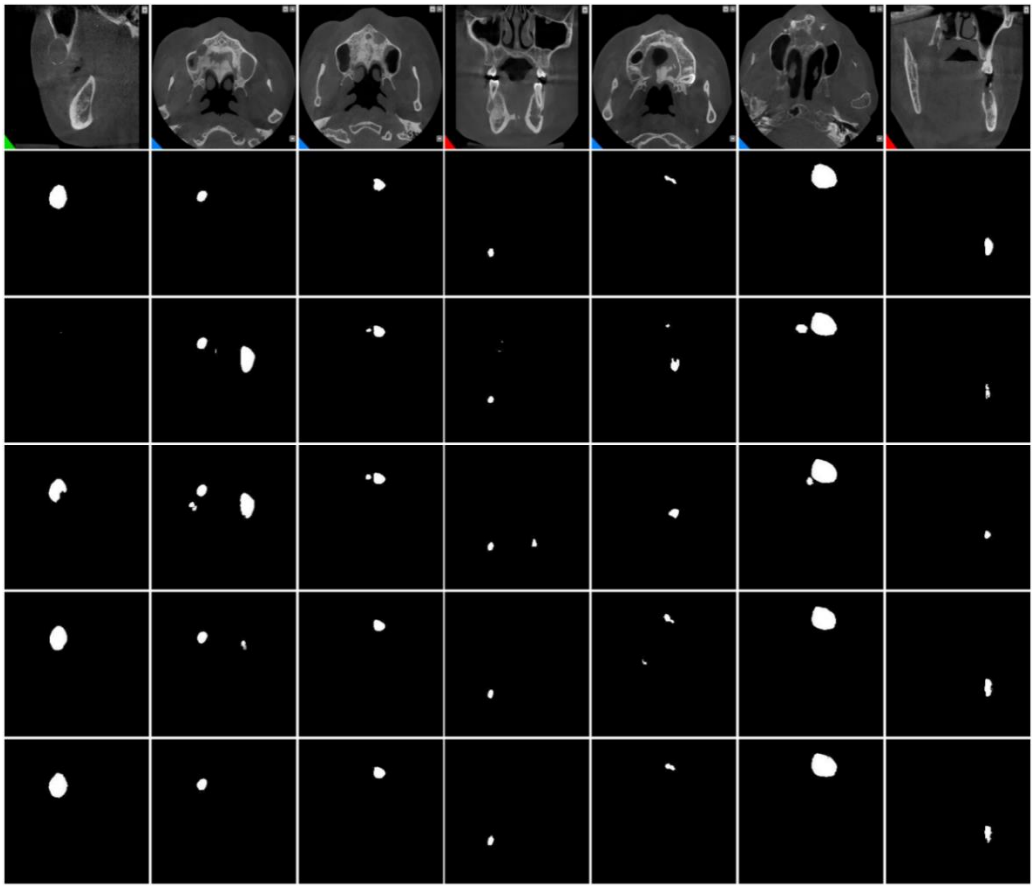

Figure 5. Ablation experiments of our method. The first and second rows show the original images with their corresponding labels. The third to last rows are the results of baseline, baseline+multiple side-outputs, baseline+feature-fusion, and baseline+multiple side-outputs+ feature-fusion.

3.5. Comparison with other models

To verify the effectiveness of the proposed approach on lesion area, thirteen algorithms such as U-Net, UNet++, UNet 3+, FSP2-Net, SAR-U-Net, FCSNet, FF-UNet, DUDA-Net, BCDU-Net, UNet++-MSOF, META-Unet, MSU-Net, and CE-Net were selected as comparison methods. Among them, UNet, UNet++, and UNet 3+ are classic segmentation algorithms, and AR-U-Net, FCSNet,

FF-UNet, BCDU-Net, and META-Unet introduce self-guided attention mechanism into the network. Additionally, the comparison involved innovative network architectures such as DUDA-Net, a double U-shaped network, FSP2-Net, and UNet++_MSOF with multiple side-outputs structure. To ensure the integrity and reliability of the comparison, the experimental settings and environmental conditions for each method were standardized. The results of these fourteen algorithms on the test set of real jaw cyst pathological images were counted in the experiment, as shown in Table 4. Among all methods, the performance of U-Net, FSP2-Net, SAR-U-Net, DUDA-Net, META-Unet, and MSU-Net is slightly different from other methods, which reflects the suggest limitations within their un-optimized encoders to effectively represent specific focal features in the segmentation process. Conversely, UNet++, UNet 3+, FCSNet, FF-UNet, BCDU-Net, and CE-Net showcased clear performance advantages, attributed to their integration of dense connections, which seemed to enhance their segmentation capabilities significantly. By introducing multiple side-outputs strategy on the basis of UNet++, UNet++-MSOF also achieves good performance improvement. However, distinctly standing out from the others, the HIMS-Net employed a unique strategy. Leveraging horizontal-vertical interaction mechanisms and feature-fusion units, the HIMS-Net effectively merged global semantic information with intricate local details. This fusion led to notably higher segmentation accuracy and a reduced missed detection rate in the segmentation of cyst lesions, highlighting its superior ability to accurately delineate cystic regions.

Table 4. Results of our method with other models.

Method	Mcc	Dice	Jaccard
U-Net [6]	0.8894	0.8891	0.8031
UNet++ [22,23]	0.8953	0.8951	0.8128
UNet 3+ [33]	0.9009	0.8999	0.8221
FSP2-Net [34]	0.8756	0.8753	0.7795
SAR-U-Net [35]	0.8490	0.8480	0.7399
FCSNet [36]	0.9106	0.9108	0.8374
FF-UNet [37]	0.8928	0.8930	0.8073
DUDA-Net [38]	0.8733	0.8722	0.7758
BCDU-Net [39]	0.8996	0.8988	0.8204
UNet++-MSOF [40]	0.9069	0.9055	0.8321
META-Unet [41]	0.8801	0.8802	0.7876
MSU-Net [42]	0.8749	0.8744	0.7803
CE-Net [43]	0.8903	0.8901	0.8036
HIMS-Net	0.9350	0.9351	0.8786

Figure 6 presents a visual comparison of the lesion area segmentation outcomes derived from the diverse methods discussed earlier. This figure serves as a comprehensive visual benchmarking tool, displaying the segmented results generated by fourteen specific algorithms: U-Net, UNet++, UNet 3+, FSP2-Net, SAR-U-Net, FCSNet, FF-UNet, DUDA-Net, BCDU-Net, UNet++-MSOF, META-Unet, MSU-Net, CE-Net, and HIMS-Net. Each of these results is placed side by side with the original image and the corresponding labels that represent the ground truth. It can be clearly seen that other models seem to lack in capturing and integrating adequate semantic information, resulting in numerous instances of missed segmentations and false detections. The discrepancy is quite evident when comparing the segmentation results of these models against the original labels. In contrast, the HIMS-Net model's performance stands out due to its ability to incorporate and combine various

modules, allowing for the effective modeling of multi-scale context information. This intricate integration empowers the network to navigate through and segregate the non-focal area information, thereby enhancing the precision and accuracy of segmenting smaller areas within the lesions.

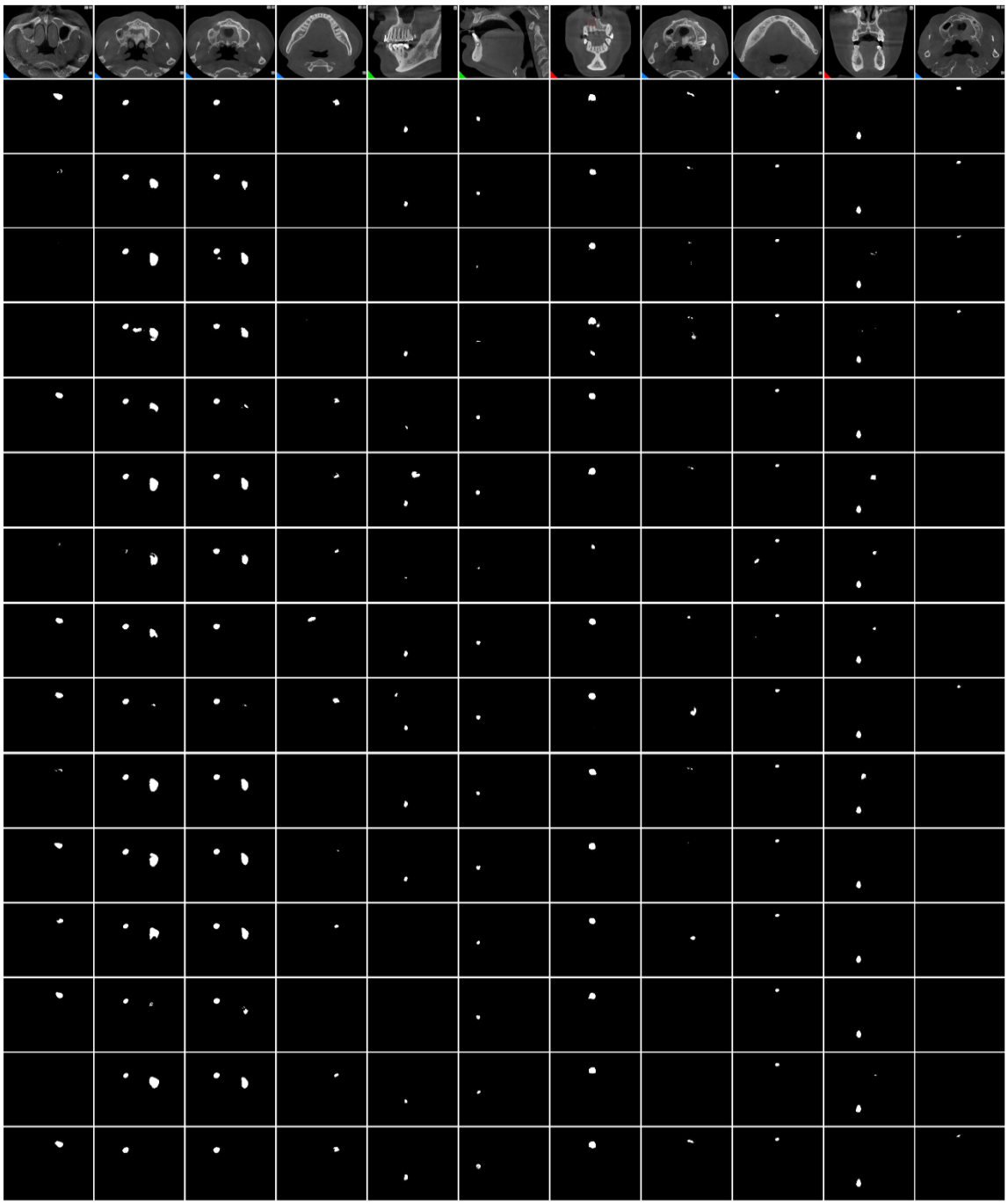

Figure 6. Visual segmentation results of various models. The first and second rows show the original images with their corresponding labels. The third to last rows are the results of U-Net, UNet++, UNet 3+, FSP2-Net, SAR-U-Net, FCSNet, FF-UNet, DUDA-Net, BCDU-Net, UNet++-MSOF, META-Unet, MSU-Net, CE-Net, and HIMS-Net.

3.6. Computational analysis and efficiency comparison

In our pursuit of fairness and reasonableness, we removed the multi-side outputs module from HIMS-Net network. This was an essential step to ensure a more equitable comparison among the models under evaluation. Upon meticulous examination presented in Table 5, it becomes evident that SAR-U-Net and U-Net distinctly demand fewer parameters in contrast to the other models. However, this reduction in parameters comes at the cost of performance. Notably, the absence of certain functionalities in SAR-U-Net and U-Net influences their overall efficacy in complex tasks. On the contrary, HIMS-Net distinguishes itself through the incorporation of a horizontal-vertical interaction mechanism and a feature-fusion unit. These critical inclusions significantly contribute to the network's prowess in understanding intricate patterns and relationships within the data. However, compared to some traditional models such as U-Net, this enhancement in capability necessitates a larger investment in both time and parameters during the training process. This is an inherent feature of HIMS-Net, attributed to its intricate architecture with the horizontal-vertical interaction mechanism and the feature-fusion unit, all of which contribute significantly to the network's capability to capture multi-scale contextual information. Despite the increased requirements, HIMS-Net manages to strike a balance between detection accuracy and computational efficiency, as evidenced by its performance within an acceptable inference time frame.

Table 5. Comparison of parameter counts and computational time among different models.

Method	Parameter (M)	Time (ms/step)
U-Net [6]	2.06	12
UNet++ [22,23]	8.62	28
UNet 3+ [33]	21.57	25
FSP2-Net [34]	74.47	96
SAR-U-Net [35]	0.51	10
FCSNet [36]	27.76	57
FF-UNet [37]	3.76	25
DUDA-Net [38]	31.10	68
BCDU-Net [39]	19.70	52
UNet++-MSOF [40]	8.62	73
META-Unet [41]	21.69	27
MSU-Net [42]	47.08	82
CE-Net [43]	29.00	61
HIMS-Net	18.24	48

3.7. Experiment on data augmentation dataset

To address the challenges posed by a limited number of samples and an imbalanced distribution within jaw cyst dataset, we adopted a range of data augmentation techniques to expand the training set and validation set. These methods included random rotation, scaling shifts, translation and clipping to ensure that our method could effectively recognize jaw cysts from different angles. After data augmentation, a total of 3991 images have been acquired. The new dataset includes 306 for testing (only containing the original jaw cyst images), 2765 for training and 920 for validation. In the experimentation phase, we conducted extensive tests using U-Net, UNet++, UNet 3+, FSP2-Net,

SAR-U-Net, FCSNet, FF-UNet, DUDA-Net, BCDU-Net, UNet++-MSOF, META-Unet, MSU-Net, and CE-Net. These models were trained and evaluated on our enhanced dataset, and the results were meticulously documented in Table 6. It can be seen from the results that the Mcc, Dice, and Jaccard values obtained by the above models are all improved after the database enhancement. Of particular interest, attributed to its architecture with the horizontal-vertical interaction mechanism, the feature-fusion unit, and multiple side-outputs, the HIMS-Net still achieves the best segmentation results among the various architectures. For a visual representation of the segmentation outcomes produced by each model are shown in Figure 7. These illustrations provide a qualitative assessment of the lesion segmentation results, further highlighting the superiority of HIMS-Net in delivering precise and visually compelling segmentations.

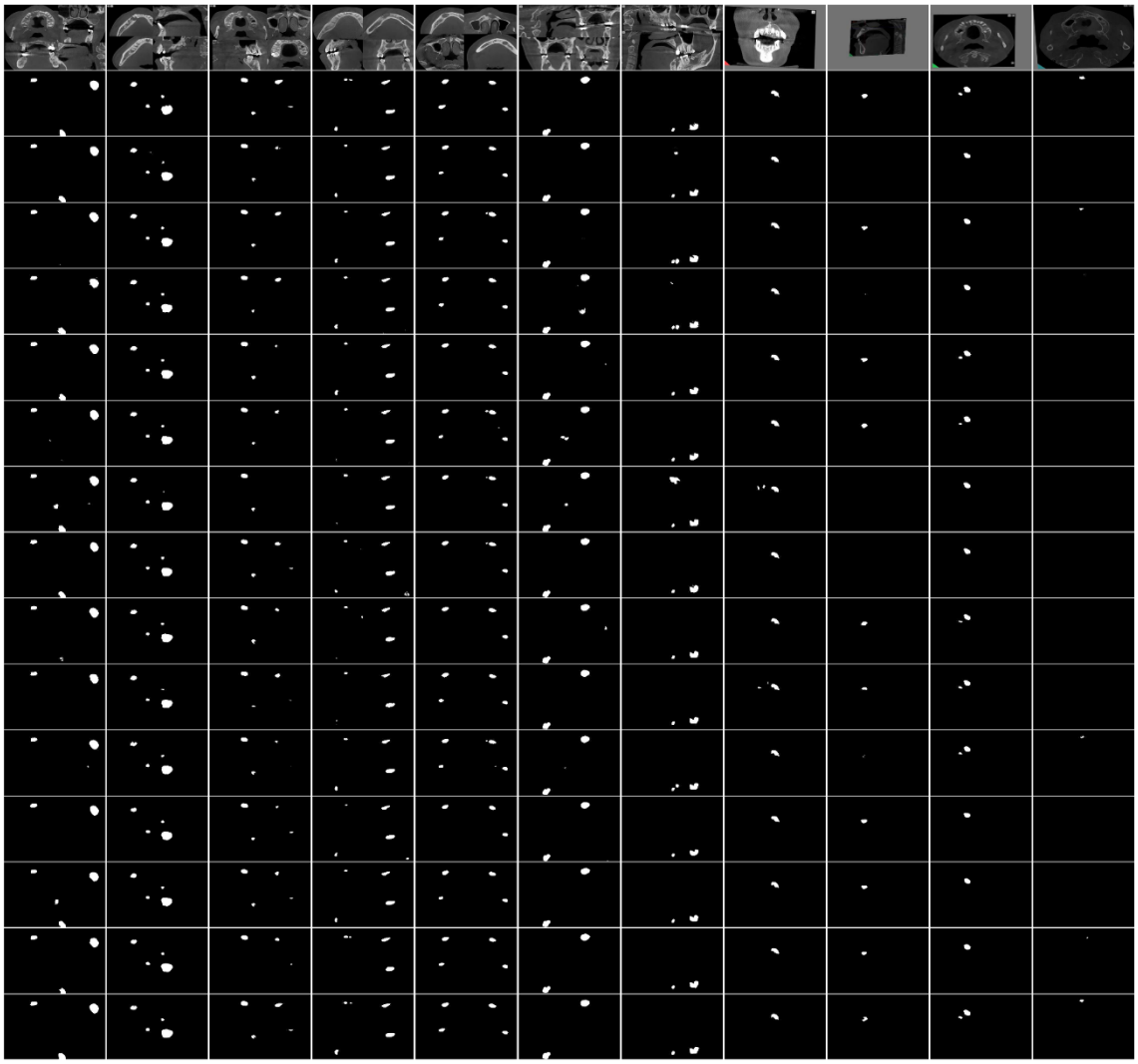

Figure 7. Visual segmentation results of various models on data augmentation dataset. The first and second rows show the original images with their corresponding labels. The third to last rows are the results of U-Net, UNet++, Unet 3+, FSP2-Net, SAR-U-Net, FCSNet, FF-Unet, DUDA-Net, BCDU-Net, Unet++-MSOF, META-Unet, MSU-Net, CE-Net, and HIMS-Net.

Table 6. Results of our method with other models on the dataset after data augmentation.

Method	Mcc	Dice	Jaccard
U-Net [6]	0.9225	0.9226	0.8576
UNet++ [22,23]	0.9073	0.9065	0.8328
UNet 3+ [33]	0.9180	0.9177	0.8502
FSP2-Net [34]	0.8970	0.8963	0.8148
SAR-U-Net [35]	0.8976	0.8971	0.8161
FCSNet [36]	0.9268	0.9272	0.8648
FF-UNet [37]	0.9184	0.9184	0.8505
DUDA-Net [38]	0.9245	0.9250	0.8613
BCDU-Net [39]	0.9186	0.9187	0.8512
UNet++-MSOF [40]	0.9159	0.9154	0.8469
META-Unet [41]	0.8975	0.8980	0.8169
MSU-Net [42]	0.8835	0.8831	0.7936
CE-Net [43]	0.9030	0.9029	0.8238
HIMS-Net	**0.9361**	**0.9366**	**0.8810**

4. Conclusions

This research is dedicated to the examination of jaw cysts, presenting an innovative deep-learning network that operates on a foundation of a novel horizontal-vertical interaction mechanism and multiple side-outputs. The relevant conclusions are as follows: First, the horizontal-vertical interaction mechanism is introduced to reduce the loss of spatial information and make the network have stronger feature reuse capability without increasing parameters. Second, a feature-fusion unit that combines extended convolution and standard convolution to obtain different receptive fields, which can have richer context information. Moreover, the multi-side outputs strategy is utilized to fuse the feature information of different semantic levels. The results show that our method achieves 93.61% of Mcc, 93.66% of Dice and 88.10% of Jaccard, which is superior to other traditional detection models. In the future, our research aims to expand the application of this innovative technology to a broader spectrum of medical image analysis tasks. With a vision to create an advanced automated diagnostic system, it is to not only differentiate but accurately predict between benign and malignant lesions.

Ethics statement

The Institutional Ethics Review Committee of Quzhou People's Hospital approved this retrospective study. All data were fully anonymized.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62102227), Zhejiang Basic Public Welfare Research Project (No. LZY24E050001, LZY24E060001), Science and Technology Major Projects of Quzhou (2022K56, 2022K92, 2023K221, 2023K211).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. P. Wang, J. Z. Peng, M. Pedersoli, Y. F. Zhou, C. M. Zhang, C. Desrosiers, CAT: Constrained adversarial training for anatomically-plausible semi-supervised segmentation, *IEEE Trans. Med. Imaging*, **42** (2023), 2146–2161. https://doi.org/10.1109/TMI.2023.3243069
2. L. Zhang, K. J. Zhang, H. W. Pan, SUNet plus plus: A deep network with channel attention for small-scale object segmentation on 3D medical images, *Tsinghua Sci. Technol*., **28** (2023), 628–638. https://doi.org/10.26599/TST.2022.9010023
3. D. D. Meng, S. Li, B. Sheng, H. Wu, S. Q. Tian, W. J. Ma, et al., 3D reconstruction-oriented fully automatic multi-modal tumor segmentation by dual attention-guided VNet, *Visual Comput.*, **39** (2023), 3183–3196. https://doi.org/10.1007/s00371-023-02965-0
4. Y. Feng, Y. H. Wang, H. H. Li, M. J. Qu, J. Z. Yang, Learning what and where to segment: A new perspective on medical image few-shot segmentation, *Med. Image Anal.*, **87** (2023), 102834. https://doi.org/10.1016/j.media.2023.102834
5. Y. X. Ma, S. Wang, Y. Hua, R. H. Ma, T. Song, Z. G. Xue, et al. Perceptual data augmentation for biomedical coronary vessel segmentation, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **20** (2023), 2494–2505. https://doi.org/10.1109/TCBB.2022.3188148
6. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, Munich, Germany, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
7. A. Sharma, P. K. Mishra, DRI-UNet: dense residual-inception UNet for nuclei identification in microscopy cell images, *Neural Comput. Appl.*, **35** (2023), 19187–19220. https://doi.org/10.1007/s00521-023-08729-0
8. B. Sarica, D. Z. Seker, B. Bayram, A dense residual U-net for multiple sclerosis lesions segmentation from multi-sequence 3D MR images, *Int. J. Med. Inf.*, **170** (2023), 104965. https://doi.org/10.1016/j.ijmedinf.2022.104965
9. Q. Xu, Z. Ma, H. E. Na, W. Duan, DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation, *Comput. Biol. Med.*, **154** (2023), 106626. https://doi.org/10.1016/j.compbiomed.2023.106626
10. H. Wang, G. Xu, X. Pan, Z. Liu, N. Tang, R. Lan, et al., Attention-inception-based U-Net for retinal vessel segmentation with advanced residual, *Comput. Electr. Eng.*, **98** (2022), 107670. https://doi.org/10.1016/j.compeleceng.2021.107670

11. J. Zhang, Y. Zhang, Y. Jin, J. Xu, X. Xu, MDU-Net: multi-scale densely connected U-Net for biomedical image segmentation, *Health Inf. Sci. Syst.*, **11** (2023), 13. https://doi.org/10.1007/s13755-022-00204-9
12. S. Banerjee, J. Lyu, Z. Huang, F. H. Leung, T. Lee, D. Yang, et al., Ultrasound spine image segmentation using multi-scale feature fusion skip-inception U-Net (SIU-Net), *Biocybern. Biomed. Eng*., **42** (2022), 341–361. https://doi.org/10.1016/j.bbe.2022.02.011
13. S. Wang, V. K. Singh, E. Cheah, X. Wang, Q. Li, S. H. Chou, et al., Stacked dilated convolutions and asymmetric architecture for U-Net-based medical image segmentation, *Comput. Biol. Med.*, **148** (2022), 105891. https://doi.org/10.1016/j.compbiomed.2022.105891
14. J. Mutaguchi, K. I. Morooka, S. Kobayashi, A. Umehara, S. Miyauchi, F. Kinoshita, et al., Artificial intelligence for segmentation of bladder tumor cystoscopic images performed by U-Net with dilated convolution, *J. Endourol*., **36** (2022), 827–834. https://doi.org/10.1089/end.2021.0483
15. J. Vidal, J. C. Vilanova, R. Martí, A U-Net ensemble for breast lesion segmentation in DCE MRI, *Comput. Biol. Med*., **140** (2022), 105093. https://doi.org/10.1016/j.compbiomed.2021.105093
16. K. Sun, Y. Xin, Y. Ma, M. Lou, Y. Qi, J. Zhu, ASU-Net: U-shape adaptive scale network for mass segmentation in mammograms, *J. Intell. Fuzzy Syst*., **42** (2022), 4205–4220. https://doi.org/10.3233/JIFS-210393
17. F. Abdolali, R. A. Zoroofi, Y. Otake, Y. Sato, Automatic segmentation of maxillofacial cysts in cone beam CT images, *Comput. Biol. Med*., **72** (2016), 108–119. https://doi.org/10.1016/j.compbiomed.2016.03.014
18. M. K. Alsmadi, A hybrid Fuzzy C-means and Neutrosophic for jaw lesions segmentation, *Ain Shams Eng. J.*, **9** (2018), 697–706. https://doi.org/10.1016/j.asej.2016.03.016
19. J. Hu, Z. Feng, Y. Mao, J. Lei, D. Yu, M. Song, A location constrained dual-branch network for reliable diagnosis of jaw tumors and cysts, in *International Conference of Medical Image Computing and Computer Assisted Intervention*, Strasbourg, France, (2021), 723–732. https://doi.org/10.1007/978-3-030-87234-2_68
20. S. Sivasundaram, C. Pandian, Performance analysis of classification and segmentation of cysts in panoramic dental images using convolutional neural network architecture, *Int. J. Imaging Syst. Technol.*, **31** (2021), 2214–2225. https://doi.org/10.1002/ima.22625
21. D. K. Veena, A. Jatti, M. J. Vidya, R. Joshi, S. Gade, A novel approach towards automatic contour identification of jaw cysts from digital panoramic radiographs to improvise the treatment planning, *Int. J. Biol. Biomed. Eng*., **16** (2022), 1–8. https://doi.org/10.46300/91011.2022.16.1
22. Z. W. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. M. Liang, UNet++: A nested U-Net architecture for medical image segmentation, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Granada, Spain, (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
23. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation, *IEEE Trans. Med. Imaging*, **39** (2020), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609

24. K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in *IEEE International Conference on Computer Vision & Pattern Recognition*, Long Beach, CA, USA, (2019), 5686–5696. https://doi.org/10.1109/CVPR.2019.00584
25. M. Zhao, Y. Wei, Y. Lu, K. K. L. Wong, A novel U-Net approach to segment the cardiac chamber in magnetic resonance images with ghost artifacts, *Comput. Methods Programs Biomed.*, **196** (2020), 105623. https://doi.org/10.1016/j.cmpb.2020.105623
26. A. S. Mahmoud, S. A. Mohamed, R. A. El-Khoriby, H. M. AbdelSalam, I. A. El-Khodary, Oil spill identification based on dual attention UNet model using synthetic aperture radar images, *J. Indian Soc. Remote Sens.*, **51** (2023), 121–133. https://doi.org/10.1007/s12524-022-01624-6
27. X. Xie, X. Pan, W. Zhang, J. An, A context hierarchical integrated network for medical image segmentation, *Comput. Electr. Eng.*, **101** (2022), 108029. https://doi.org/10.1016/j.compeleceng.2022.108029
28. L. Zhu, L. Zhang, W. Hu, H. Chen, H. Li, S. Wei, et al., A multi-task two-path deep learning system for predicting the invasiveness of craniopharyngioma, *Comput. Meth. Prog. Bio.*, **216** (2022), 106651. https://doi.org/10.1016/j.cmpb.2022.106651
29. L. Zhang, Y. Liao, G. Wang, J. Chen, H. Wang, A multi-scale contextual information enhancement network for crack segmentation, *Appl. Sci.*, **12** (2022), 11135. https://doi.org/10.3390/app122111135
30. M. Jiang, X. Zhang, Y. Sun, W. Feng, Q. Gan, Y. Ruan, AFSNet: Attention-guided full-scale feature aggregation network for highresolution remote sensing image change detection, *GISci. Remote Sens.*, **59** (2022), 1882–1900. https://doi.org/10.1080/15481603.2022.2142626
31. C. Xu, Y. Qi, Y. Wang, M. Lou, J. Pi, Y. Ma, ARF-Net: An adaptive receptive field network for breast masssegmentation in whole mammograms and ultrasound images, *Biomed. Signal Process. Control*, **71** (2022), 103178. https://doi.org/10.1016/j.bspc.2021.103178
32. D. Maji, P. Sigedar, M. Singh, Attention Res-UNet with guided decoder for semantic segmentation of brain tumors, *Biomed. Signal Process. Control*, **71** (2022), 103077. https://doi.org/10.1016/j.bspc.2021.103077
33. H. Huang, L. Lin, R. Tong, H. Hu, J. Wu, UNet 3+: A full-scale connected UNet for medical image segmentation, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, (2020), 1055–1059. https://doi.org/10.1109/ICASSP40776.2020.9053405
34. K. Wang, S. Liang, S. Zhong, Q. Feng, Y. Zhang, Breast ultrasound image segmentation: A coarse-to-fine fusion convolutional neural network, *Med. Phys.*, **48** (2021), 4262–4278. https://doi.org/10.1002/mp.15006
35. J. Wang, P. Lv, H. Wang, C. Shi, SAR-U-Net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver ct segmentation, *Comput. Methods Programs Biomed.*, **208** (2021), 106268. https://doi.org/10.1016/j.cmpb.2021.106268
36. G. Xiao, B. Zhu, Y. Zhang, H. Gao, FCSNet: A quantitative explanation method for surface scratch defects during belt grinding based on deep learning, *Comput. Ind.*, **144** (2023), 103793. https://doi.org/10.1016/j.compind.2022.103793
37. A. Iqbal, M. Sharif, M. A. Khan, W. Nisar, M. Alhaisoni, FF-UNet: A u-shaped deep convolutional neural network for multimodal biomedical image segmentation, *Cognit. Comput.*, **14** (2022), 1287–1302. https://doi.org/10.1007/s12559-022-10038-y

38. F. Xie, Z. Huang, Z. Shi, T. Wang, G. Song, B. Wang, et al., DUDA-Net: A double u-shaped dilated attention network for automatic infection area segmentation in COVID-19 lung CT images, *Int. J. Comput. Assisted Radiol. Surg.*, **16** (2021), 1425–1434. https://doi.org/10.1007/s11548-021-02418-w
39. R. Azad, M. Asadi-Aghbolaghi, M. Fathy, S. Escalera, Bi-directional ConvLSTM U-Net with densley connected convolutions, in *Proceedings of the IEEE/CVF International Conference on Computer Vision workshops*, Seoul, Korea, (2019), 406–415. https://doi.org/10.1109/ICCVW.2019.00052
40. D. Peng, Y. Zhang, H. Guan, End-to-end change detection for high resolution satellite images using improved UNet++, *Remote Sens.*, **11** (2019), 1382. https://doi.org/10.3390/rs11111382
41. H. Wu, Z. Zhao, Z. Wang, META-Unet: Multi-scale efficient transformer attention Unet for fast and high-accuracy polyp segmentation, *IEEE Trans. Autom. Sci. Eng.*, (2023), 1–12. https://doi.org/10.1109/TASE.2023.3292373
42. R. Su, D. Zhang, J. Liu, C. Cheng, MSU-Net: Multi-scale U-Net for 2D medical image segmentation, *Front. Genet*., **12** (2021), 639930. https://doi.org/10.3389/fgene.2021.639930
43. Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, et al., CE-Net: Context encoder network for 2d medical image segmentation, *IEEE Trans. Med. Imaging*, **38** (2019), 2281–2292. https://doi.org/10.1109/TMI.2019.2903562
44. M. M. Ji, Z. B. Wu, Automatic detection and severity analysis of grape black measles disease based on deep learning and fuzzy logic, *Comput. Electron. Agric.*, **193** (2022), 106718. https://doi.org/10.1016/j.compag.2022.106718
45. M. Jiang, F. Zhai, J. Kong, A novel deep learning model DDU-net using edge features to enhance brain tumor segmentation on MR images, *Artif. Intell. Med.*, **121** (2021), 102180. https://doi.org/10.1016/j.artmed.2021.102180
46. Y. Y. Yang, C. Feng, R. F. Wang, Automatic segmentation model combining U-Net and level set method for medical images, *Expert Syst. Appl.*, **153** (2020), 113419. https://doi.org/10.1016/j.eswa.2020.113419
47. C. Zhao, R. J. Shuai, L. Ma, W. J. Liu, M. L. Wu, Segmentation of dermoscopy images based on deformable 3D convolution and ResU-NeXt++, *Med. Biol. Eng. Comput.*, **59** (2021), 1815–1832. https://doi.org/10.1007/s11517-021-02397-9

AIMS Press