scientific reports

OPEN

A high-order focus interaction model and oral ulcer dataset for oral ulcer segmentation

Chenghao Jiang[1,7], Renkai Wu[2,7], Yinghao Liu[6], Yue Wang[2], Qing Chang[2], Pengchen Liang[2]✉ & Yuan Fan[3,4,5]✉

Computer-aided diagnosis has been slow to develop in the field of oral ulcers. One of the major reasons for this is the lack of publicly available datasets. However, oral ulcers have cancerous lesions and their mortality rate is high. The ability to recognize oral ulcers at an early stage in a timely and effective manner is a very critical issue. In recent years, although there exists a small group of researchers working on these, the datasets are private. Therefore to address this challenge, in this paper a multi-tasking oral ulcer dataset (Autooral) containing two major tasks of lesion segmentation and classification is proposed and made publicly available. To the best of our knowledge, we are the first team to make publicly available an oral ulcer dataset with multi-tasking. In addition, we propose a novel modeling framework, HF-UNet, for segmenting oral ulcer lesion regions. Specifically, the proposed high-order focus interaction module (HFblock) performs acquisition of global properties and focus for acquisition of local properties through high-order attention. The proposed lesion localization module (LL-M) employs a novel hybrid sobel filter, which improves the recognition of ulcer edges. Experimental results on the proposed Autooral dataset show that our proposed HF-UNet segmentation of oral ulcers achieves a DSC value of about 0.80 and the inference memory occupies only 2029 MB. The proposed method guarantees a low running load while maintaining a high-performance segmentation capability. The proposed Autooral dataset and code are available from https://github.com/wurenkai/HF-UNet-and-Autooral-dataset.

Keywords Computer-aided diagnosis, Medical image segmentation, Oral ulcer, Deep learning, High-order interactions

Computer vision and assisted analysis have been widely used in the field of medical images. Specifically, medical image segmentation models can effectively assist doctors in diagnosis. In the field of dentistry, oral ulcers are characterized by persistent destruction of the integrity of the oral epithelium. At the same time, there is a variable loss of the underlying connective tissue with a pothole-like appearance[1]. Cancerous oral ulcers also exist, and oral cancer is more prevalent in people over 40 years of age, and specifically, the incidence in men tends to be twice as high as the incidence in women[2,3]. Oral cancer is characterized by high mortality, late detection and high morbidity[4]. The predisposing forms of oral cancer are closely related to any form of smoking and heavy alcohol consumption. However, during the diagnostic process performed by the dentist, the low contrast of the diseased area and the small size of the area lead to leakage and misdiagnosis[5–7]. And one of the important methods to reduce the doctor's missed diagnosis is computer-aided diagnosis (CAD).

Computer-aided diagnosis utilizes computer vision and artificial intelligence to automatically detect, segment and identify lesions in medical images. This can improve the efficiency of a doctor's diagnosis. In particular, for the problem of uneven distribution of medical resources across geographic regions, the technology can reduce this gap. Currently deep learning algorithms are mainly used for medical image segmentation. Full convolutional neural network (FCN)[8] is a pioneer in image segmentation, where features are extracted by convolution. In 2015, UNet model[9] was designed on the basis of FCN. The emergence of UNet model has taken CAD a step

[1]Stomatological College, Nanjing Medical University, Nanjing, China. [2]School of Microelectronics, Shanghai University, Shanghai, China. [3]Department of Oral Mucosal Diseases, The Affiliated Stomatological Hospital of Nanjing Medical University, Nanjing, China. [4]Jiangsu Province Key Laboratory of Oral Diseases, Nanjing Medical University, Nanjing, China. [5]Jiangsu Province Engineering Research Center of Stomatological Translational Medicine, Nanjing, China. [6]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China. [7]These authors contributed equally: Chenghao Jiang and Renkai Wu. ✉email: liangpengchen@shu.edu.cn; fanyuan@njmu.edu.cn

further. Until now, there are still many researchers who have proposed models with better performance based on the UNet model. In this paper, one of the research results is to propose a better performance model based on the UNet model.

In Yang et al.[10], researchers proposed a novel mechanism of focal self-attention. It combines fine-grained local and coarse-grained global interactions, which results in the formation of focal transformers with both short-term and long-term dependencies. The focal self-attention mechanism is able to have a smaller time cost and a larger receptive field than the traditional standard self-attention mechanism[11]. In Yang et al.[12], a new model structure using focal self-attention mechanism was proposed by Microsoft researchers. The proposed model outperforms the state-of-the-art models of the time in image classification, target detection and image segmentation. In this study, we take a step higher in the focal self-attention mechanism. We propose a model with high-order focus interaction (HF-UNet). However, traditional focal self-attention only realizes second-order spatial interactions at the same scale using focal transformers[13]. In particular, researchers[13,14] have proposed that realizing higher-order spatial interactions at the same scale can significantly improve feature learning. In this study, we address the lack of higher-order interactions in traditional focal self-attention mechanisms. Specifically, we take a step higher on the focal self-attention mechanism and propose a model with higher-order focal interactions (HF-UNet) for oral ulcer segmentation. This is because oral ulcers have similar feature information by interfering with teeth and reflections, and it is further demonstrated through experiments that HF-UNet can solve this problem well.

The lack of oral ulcer datasets is a major problem that hinders the development of computer-aided diagnosis in the field of dentistry. Even though some researchers have made studies in the field[15–17], the datasets are still private and not made public. To the best of our knowledge, there are no publicly available high-quality oral ulcer datasets, and this important reason is due to privacy concerns. Because the oral region is located in the face position, it is easy to disclose patient information if the image is not processed properly. We process the data in detail to ensure that patient information is not compromised. To the best of our knowledge, we are the first team to make publicly available a multi-tasking oral ulcer dataset, named Autooral dataset, which is of great significance for the entry of CAD into the field of oral ulcers. Oral ulcers with low contrast and small lesion areas (as shown in Fig. 1) is an urgent problem that needs to be solved.

In this paper, we have conducted a series of studies based on our proposed oral ulcer dataset (Autooral dataset). We propose a high-order focus interaction model (HF-UNet) on the model, which can well solve the problems of low contrast and small lesion area. Comparing with the current state-of-the-art medical image segmentation models, we obtain the best performance. More specifically, we also propose a lesion localization model to further address the aforementioned oral ulcer problem. The contributions of this paper can be summarized as follows:

- A multi-tasking oral ulcer dataset (Autooral dataset) is proposed. To the best of our knowledge, we are the first team to make publicly available a multi-tasking oral ulcer dataset. We have addressed patient privacy issues accordingly and have been approved for public availability by the appropriate organizations.
- A novel model architecture, called HF-UNet, is proposed for medical image segmentation of oral ulcers. Specifically, we propose a high-order focus interaction module (HFblock) and a lesion localization module

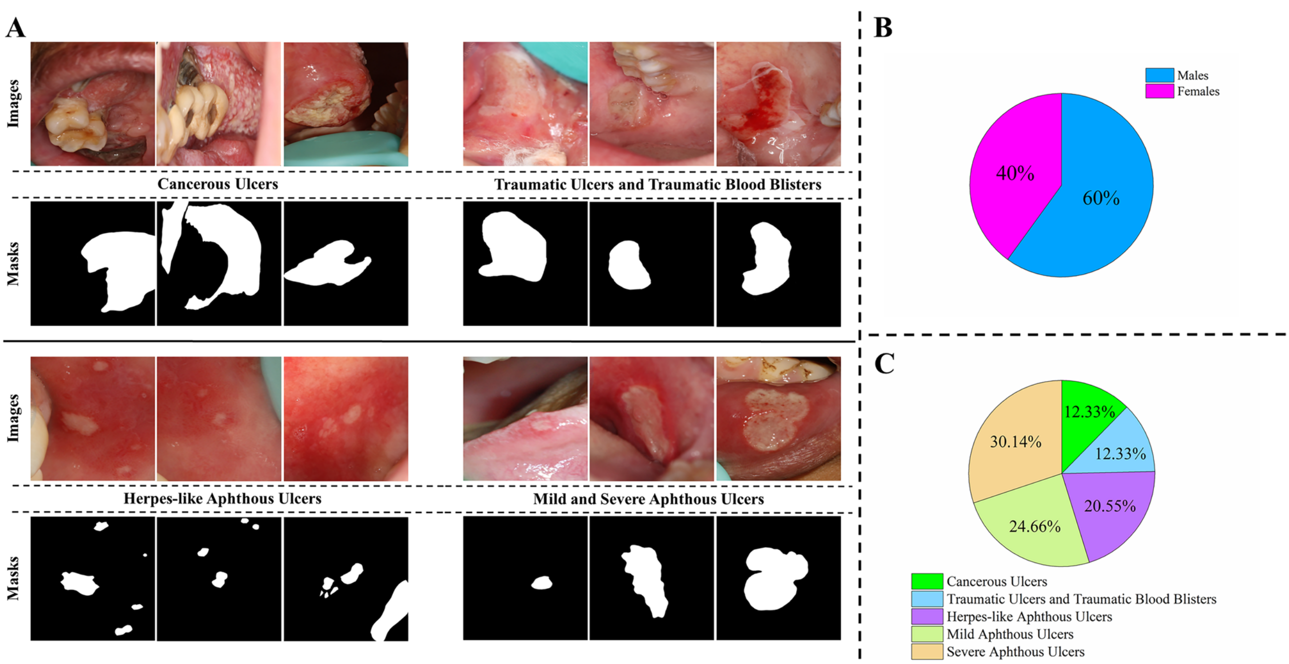

Figure 1. (**A**) Some examples from the proposed Autooral dataset. (**B**) Demographics of the proposed Autooral dataset. (**C**) The percentage of different disease categories for the proposed Autooral dataset.

(LL-M). And additional ablation experiments were performed to validate the effectiveness of the different design choices.

- The proposed model architecture outperforms the state-of-the-art methods on Autooral dataset compared to other medical image segmentation methods.
- We will make the link to the Autooral dataset and the project code publicly available on GitHub.

Related work

Image segmentation

Image segmentation is a major key problem in computer vision. At present, deep learning is the most important technology for image segmentation, which is carried out by classifying pixel by pixel, which is essentially a classification problem. Traditional image segmentation methods are fast, low computational complexity, and may be a choice in areas where segmentation accuracy is not required. However, in the field of medical image segmentation, which requires high segmentation accuracy, the traditional methods cannot meet the practical needs. The emergence of full convolutional networks (FCN) has made image segmentation based on deep learning methods occupy a major position. At the same time, the emergence of UNet[9] has led to the rapid development of medical image segmentation based on deep learning, and the skip-connection part of UNet can make it possible to fuse the low-level and high-level features, which is very crucial for medical image segmentation with high detail requirements.

Medical image segmentation

Many of the current medical image segmentations are based on UNet with improvements. Att U-net[18] adds the attention gate notation to the original UNet to suppress irrelevant feature information in the image. TransNorm[19] is a model with few parameters, which utilizes both CNNs and Transformers to improve the generalization ability of the model. TransNorm applies the attention mechanism in the encoding and skip-connection part to adaptively calibrate the feature expression ability. MALUNet[20] is a model with very few parameters, which utilizes spatial attention maps obtained by fusing multi-stage and multi-scale feature information in the UNet, which allows the model to achieve better performance with only a low number of parameters. In Ullah et al.[21], researchers proposed a multiscale residual attention UNet for segmenting brain tumors in MRI images. It utilizes a cascade approach for multi-scale learning and adaptively learns and segments brain tumor regions. In the past year, researchers have been proposing more and more models with better performance. M^2SNet[22] is to replace the skip-connection part of the original UNet with the connectivity part of the decoding part by making subtractive connections. The traditional splicing and element-by-element summing brings a lot of redundant information and reduces the segmentation accuracy. In Ullah et al.[23], researchers propose a Dense Attention Mechanism Network (DAM-Net) for the automatic detection of COVID in chest X-rays. DAM-Net proposes to employ the use of a channel attention approach to adaptively establish the weights of individual feature channels to reduce the introduction of redundant features. C^2SDG[24] in order to improve the generalization ability of the model, it is proposed to use the shallow features of each image as well as the augmented corresponding features for comparison training and get the best performance in each comparison model. Recently, the very popular Segment Anything Model (SAM)[25] has gained a lot of attention. Many researchers have also tried to use it in the field of medical image segmentation, among which MSA[26] has achieved better performance, which has fine-tuned the SAM model in the field of medical image segmentation through the Adapter module.

In this paper, we propose a novel modeling architecture. Based on focus attention, a high-order focus interaction model (HF-UNet) is proposed for oral ulcer segmentation. Also, we propose a plug-and-play lesion localization module to improve the recognition of the contours of oral ulcers. The specific modeling approach will be elaborated in the next section.

Methods

Proposed dataset

Deep learning needs to be driven by accurate data. In particular, in the medical field, data becomes even more precious. Medical images are often mishandled to reveal patient information. We go through a series of processing (clipping and removal, etc.) to ensure the public legitimacy of our data. In this paper, we propose the Autooral dataset. This is the first publicly available multi-tasking oral ulcer dataset. Autooral dataset contains two major tasks of disease segmentation and classification. This is summarized in Table 1.

Our proposed Autooral dataset was collected from the Affiliated Stomatological Hospital of Nanjing Medical University. The study was approved by the Ethical Review Committee of the Affiliated Stomatological Hospital

Proposed dataset	Source	Total number of samples	Age range	Disease category	Percentage
Autooral	Ruijin Hospital, Shanghai Jiao Tong University School of Medicine	420	7 to 84	Cancerous ulcers	12.33%
				Traumatic ulcers and traumatic blood blister	12.33%
				Herpes-likeaphthous ulcers	20.55%
				Mild aphthous ulcers	24.66%
				Severe aphthous ulcers	30.14%

Table 1. Overview of the proposed Autooral dataset.

of Nanjing Medical University. All authors unanimously affirm that the relevant provisions of the approval were strictly adhered to in the data used and in the experimental protocol. In addition, all authors confirmed that informed consent was obtained from all subjects and/or their legal guardians. For data privacy processing, we follow strict rules to remove all content containing personal information, which includes potentially compromising image data as well as file naming. All processed data has been checked by all authors. We confirm that all methods were performed in accordance with the relevant guidelines and regulations.

Specifically, in order to increase the diversity and representativeness of the Autooral dataset, we collected cases with almost full age coverage. The Autooral dataset also incorporates cases containing a wide range of underlying diseases at the same time. What's more, the proposed dataset covers different types, degrees, and stages of lesions, which include cancerous ulcers, traumatic ulcers and traumatic blood blisters, herpes-like aphthous ulcers, mild aphthous ulcers, and severe aphthous ulcers. In addition, in order to increase the reliability of the Autooral dataset, the annotation of our dataset was performed by three dentists with extensive clinical experience.

Autooral dataset collected 80 clinical cases. We obtained a total of 420 images after a series of pre-processing. More specifically, the proposed Autooral dataset was collected from 2010 to 2023, with an average patient age of 49 years (range from 7 to 84, which almost covers all ages) and a male-female ratio of 3:2. The patient population included no underlying disease, and the presence of 12 underlying diseases such as anemia, hypertension, and nasopharyngeal cancer. For the segmentation task, as verified by our experiments, our proposed HF-UNet model obtains a DSC value of around 0.80 when tested on unseen cases. In comparison with other current state-of-the-art models, HF-UNet achieves the best performance.

Medical resources are precious, and medical data with high-quality annotation are even more precious. The labeling of the Autooral dataset was done by three experienced dentists. At the end of the annotation, we formed 420 images of oral data with high quality (as shown in Fig. 1) after cropping and removal operations. In particular, as can be seen from Fig. 1, the cropped images retained information only in small areas with ulcerative lesions, which is an important means of effectively avoiding leakage and regeneration of sensitive patient information. In addition, the data were scrutinized by all authors. We standardize the image size to 256×256. The original image is a 24-bit RGB image, the ground truth for the segmentation task is an 8-bit image, and there are five different disease types for the classification task (including cancerous ulcers, traumatic ulcers and traumatic blood blisters, herpes-like aphthous ulcers, mild aphthous ulcers, severe aphthous ulcers). The ratio of the 5 different ulcer types for the classification task was 9:9:15:18:22 (with a few exclusions). Further, by chi-square test, there were significant differences in gender (p=0.04) and age (p=0.01) of the patients among the 5 ulcer types. All-age coverage, a 13-year collection interval, and the presence of 12 underlying diseases, among other things, demonstrated that we had sufficient sample diversity.

Overall model architecture

Our proposed HF-UNet is a segmentation model framework with high-order focus interactions, which is mainly used for lesion segmentation of oral ulcers. The overall model framework diagram is shown in Fig. 2. It can be summarized that (1) firstly, the image information of an oral ulcer with the size of 256×256 and the number of channels is 3 is fed into the model framework, and undergoes one ordinary convolutions with convolution kernel of 3 at stages 1-3 respectively. (2) At stages 4-6, each stage has a high-order focus interaction module (HFblock), a lesion localization module (LL-M), and an ordinary convolution, respectively. Each stage first goes through the HFblock for extracting and fusing the focus information of different orders. Then the edge and shape feature information is extracted by localizing the lesion location through LL-M. (3) In the skip-connection path, we design the multi-dilated attention gate (MDAG), which suppresses the unimportant features and highlights the useful feature information. (4) We maintain the symmetry of the UNet structure and set the encoder to be consistent with the decoder. The decoder receives the fused feature-enhanced information enhanced by MDAG and the information from the encoder, and the feature-enhanced information provides additional complementary information to the decoder. In the following we will elaborate on our proposed module.

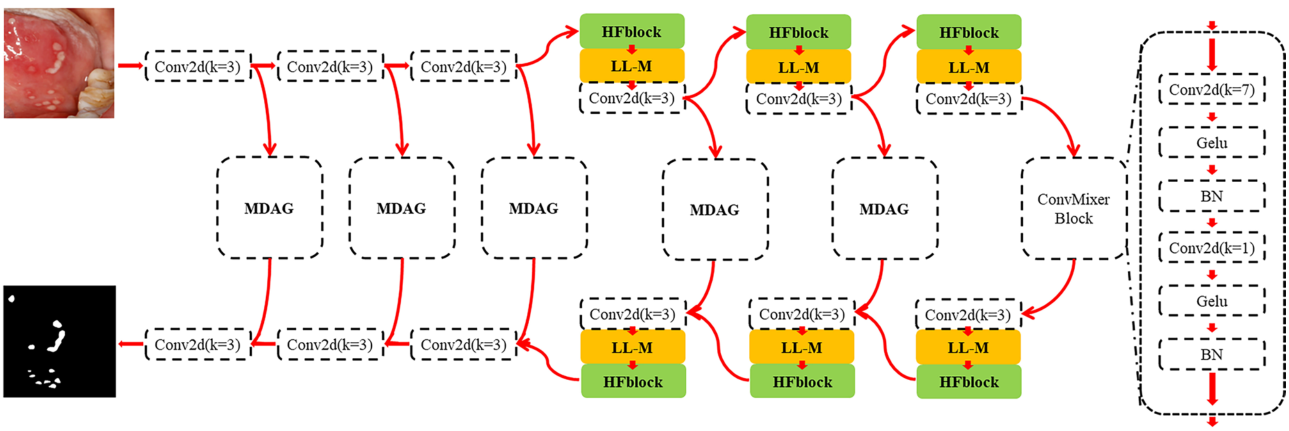

Figure 2. The architecture overview of our HF-UNet.

High-order focus interaction module

As shown in Fig. 3a, the high-order focus interaction module (HFblock) uses a design similar to the module of Transformers, with the self-attention layer inside the module replaced by a high-order focus convolution (HFConv), and a Dropout layer is used in residual linking to improve the generalization of the model. In the following part of the subsection, we introduce an efficient operation HFConv that realizes long-term, focal attention and high-order spatial interactions, as shown in Fig. 3b. HFConv consists of a linear projection, a focus module, an attention gate, and a global-local filter.

The quadratic complexity of the self-attention input in the vision Transformers shows a very rapid increase in complexity with the higher resolution requirements of the feature map. Therefore, we employ a focus module, global-local filters, and other operations to perform high-order spatial interactions, rather than reducing the complexity of self-attention as is currently employed in other literature[27,28].

First-order focus interactions

HFConv-based operations are a key part of realizing long-term, focal attention and high-order spatial interactions. For clarity, we proceed from the first-order focus interaction operation (1FConv), where $x \in \mathbb{R}^{HW\times C}$ is the input feature and the output via 1FConv can be written as:

$$\left[A_0^{HW\times C}, B_0^{HW\times C}\right] = Pro_{in}(x) \in \mathbb{R}^{HW\times 2C}, \tag{1}$$

$$P = \{AG[GLF(B_0), F(A_0)]\} \in \mathbb{R}^{HW\times C}, \tag{2}$$

$$y = Pro_{out}(P) \in \mathbb{R}^{HW\times C}, \tag{3}$$

where *Pro* is the linear projection layer, the input feature x passes through the linear projection layer, the number of channels is converted from C to $2C$ and is assigned to A_0, B_0, respectively, and the number of A_0 and B_0 channels is C. *GLF* performs the global-local filter (GLF), F performs the focus module (FM), and *AG* performs the attentional gate (AG). The first-order focus interaction between neighboring features A_0 and B_0 can be introduced by the attention gate and linear projection layer.

Global-local filter (GLF) The global-local filter (GLF) composition structure is shown in Fig. 3c. In Rao et al.[29], a global filter (GF) is proposed which has the ability to multiply frequency domain features with a learnable global filter. We adapt the GF by passing the input through the Layer Norm and then performing the GF (global) and performing two ordinary convolution operations (local), the two convolutions are 1×1 convolution, and 3×3 convolution, respectively. The 1×1 convolution reduces the complexity of the 3×3 convolution operation by reducing the number of channels by half. Finally the two operations are concatenated and output through only one Layer Norm layer to retain more feature information. Global Filtering (GF) employs a 2D Discrete Fourier Transform (2D FFT), elementwise multiplication of frequency domain features and a learnable global filter, and a 2D Inverse Fourier Transform (2D IFFT), in place of the self-concerned layer in the vision changer. The basic idea of the GF lies in its ability to cover all frequencies and to learn interactions between spatial locations in the frequency domain. This gives GF the ability to capture both long-term and short-term interactions. Unlike the GLF proposed by Ref.[13], we are using full channels for global and local operations respectively.

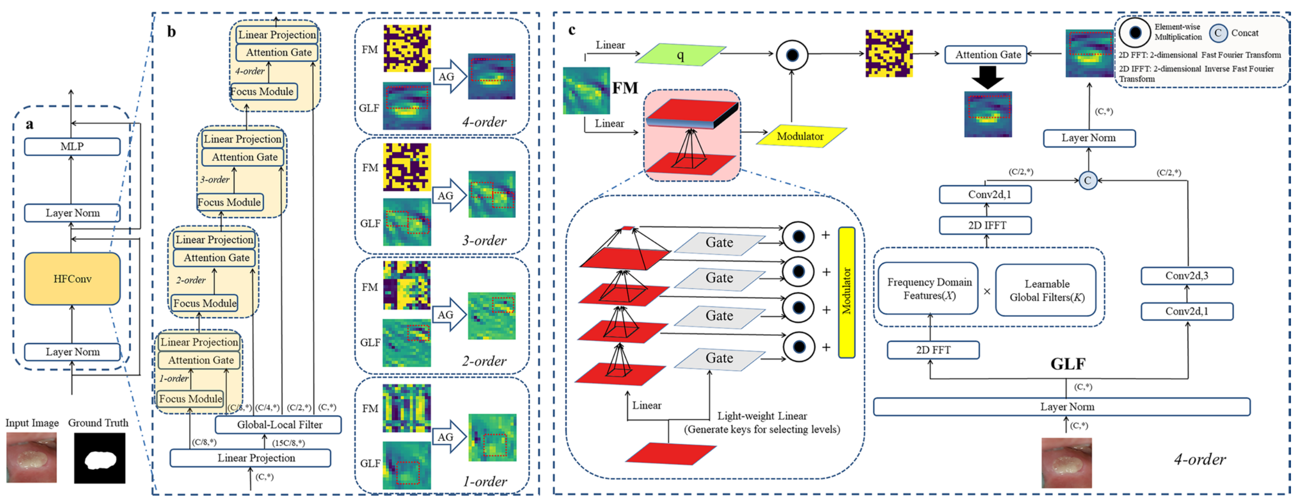

Figure 3. (**a**) Compositional structure of the proposed high-order focus interaction module. (**b**) Compositional structure of the proposed HFConv. The example shows a schematic diagram of 4FConv. FM means Focus module. GLF means Global-Local Filter. The visualization is the output of the individual modules in the last HFblock of the decoder. (**c**) Perform an interactive operation (4-order). The Gate in the FM decides what level of local fine-grained features to output.

Focus module (FM) Focus modules (FM) begin with focus transformers[10], which utilize the focus self-attention mechanism. The FM, as in Fig. 3c, is its constituent structure. The focal self-attention mechanism has a larger receptive domain than the traditional standard self-attention mechanism, while the focal self-attention mechanism also has the ability to utilize short-term dependencies and long-term dependencies, and a Dropout layer is used in residual linking to improve the generalization of the model. In Yang et al.[12], a focus network formed using the focal self-attention mechanism instead of the standard self-attention mechanism is proposed. In Naderi et al.[11], it is proposed to form Focal-UNet by combining the focal module with the UNet model framework. In this paper, we make use of the focal module to novelly propose a high-order focus module. As in Fig. 3b, we lead the focus module to higher order interactions for more comprehensive and detailed feature learning.

Attention gate (AG) The attention gate (AG) suppresses irrelevant background information, similar to Att-UNet[18], and the gating coefficients are obtained by additive attention. Additive attention has better performance than the traditional multiplicative attention performance. The AG operates as follows Eq:

$$AG_1(g,x) = Relu\{BN[Conv(g)] + BN[Conv(x)]\}, \tag{4}$$

$$AG_2(g,x) = Sig\{BN[Conv(AG_1(g,x))]\}, \tag{5}$$

$$AG(g,x) = x \cdot AG_2(g,x), \tag{6}$$

where $Conv$ denotes the convolution operation with a convolution kernel of 1, BN denotes the batch normalization operation, $Relu$ denotes the Relu activation function, Sig denotes the Sigmoid activation function, g denotes the output after each order of the linear projection layer, and x denotes the output after the global-local filter.

High-order focus interactions

By detailing the first-order focus interaction operation (1FConv), we can generalize the first-order to the n-order (n denotes any order) by forming high-order focus interactions in order to improve the extraction of features by the model. The n-order focus interactions whereas first need to pass through the linear projection layer, forming a set of A_0 and $\{B_k\}_{k=0}^{n-1}$. The specific form is given in the following equation:

$$\left[A_0^{HW\times C_0}, \cdots, B_{n-1}^{HW\times C_{n-1}}\right] = Pro_{in}(x) \in \mathbb{R}^{HW\times 2C}, \tag{7}$$

High-order focus interactions are formed by continuously performing operations and the output is scaled down by $1/\alpha$ for stabilizing model training:

$$A_{k+1} = Pro\{AG[GLF(B_k), F(A_k)]\}/\alpha, \tag{8}$$

From the above equation, it can be seen that each time an operation is performed, k is increased by 1. By continuously performing this operation in order to realize the n-order focus spatial interaction, it can be seen through Fig. 3b that the number of channels given from the global-local filter increases as the interaction order gets higher. This is a coarse-to-fine approach to spatial interaction, and the exact number of channels given at each order can be derived using the following equation:

$$C_k = \frac{C}{2^{n-k-1}}, 0 \le k \le n-1, \tag{9}$$

Through the above operation, we led the traditional focal mechanism to a higher-order space for interaction, which further improved the feature extraction ability of the focal mechanism. Specifically, it can be learned from Fig. 7b of the ablation experiments that the performance is optimal when keeping the focus mechanism at the 4-order interaction. Lower spatial interactions do not allow for good learning of feature information. However, too high spatial interactions make the number of input focus mechanism channels too sparse at the first order (Eq. 9), and keeping the appropriate order spatial interactions can achieve the best performance. Through the above analysis and ablation results, it can be more intuitively concluded that the high-order focus interaction operation proposed in this study can significantly improve the feature extraction capability of the traditional focus mechanism.

Lesion localization module

In the process of lesion segmentation by the model, contour recognition is a key element to improve the accurate segmentation of lesions. The accurate outlining of contours requires the model to have a good extraction capability of the lesion edge information. For the extraction of boundary information, the sobel operator is commonly used to obtain the gradient map[30,31]. Unlike the previous sobel operator which is only used in two and four directions, we design an eight-direction lesion localization module using a hybrid 3×3 and 5×5 sobel filter composition as shown in Fig. 4A.

As shown in Fig. 4B, the lesion localization module consists of an ordinary convolution and a filter module. The input information is first convolved with a convolution kernel of 1 and 3 and the number of channels is reduced to half and input to the filter module. The purpose of using two convolutions and then inputting them to the filter is to carry out the extraction of feature information and halving the number of channels helps to reduce the computational complexity of the next filter module. The filter module consists of eight main direction-specific sobel filters. We make each channel of the input feature information perform the 8 direction-specific sobel filters once, and then finally superimpose them to form a unified feature information. The feature information output

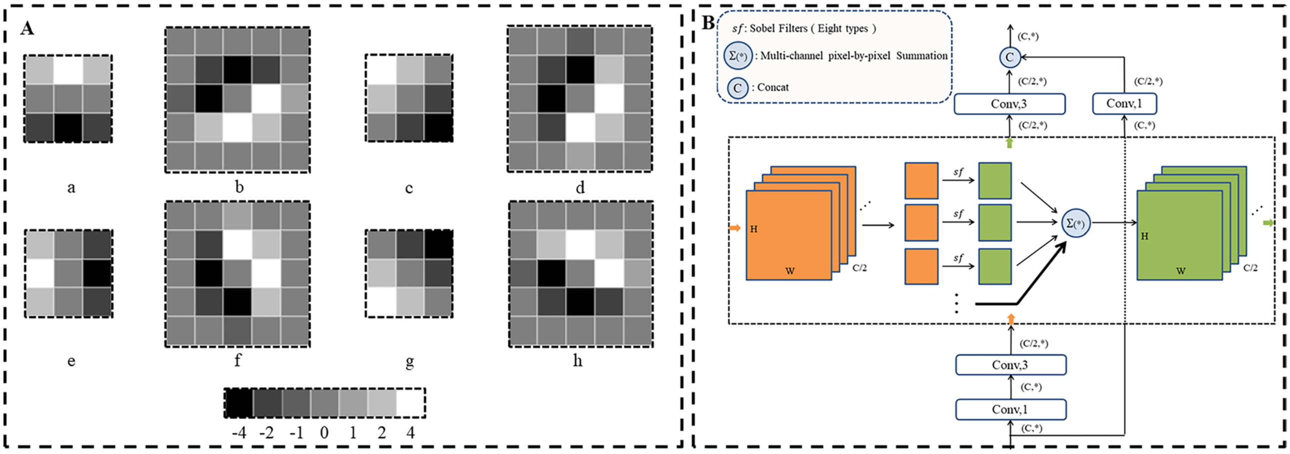

Figure 4. (**A**) Schematic of the proposed hybrid multiscale eight-direction sobel filter. (**B**) The architecture overview of our Lesion localization module.

from the filter module is then 3×3 convolved and concatenated with the original input feature information. The specific realization steps can be expressed by the following equation:

$$F_1 = Conv_{3\times3}[Conv_{1\times1}(x)], \tag{10}$$

$$F_2 = \sum_{k=1}^{8} \mathcal{Q}\ (S_k, F_1), \tag{11}$$

$$Out = Concat[Conv_{3\times3}(F_2), Conv_{1\times1}(x)], \tag{12}$$

where $\mathcal{Q}$ denotes the convolution operator and S_k is the kth sobel filter, where the eight-direction sobel operator abcdefgh denotes the directions 0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5° respectively. Each of the eight directional sobel filters emphasizes the edge features in the corresponding direction, and then a uniform feature map is generated by pixel-by-pixel summation.

Multi-dilated attention gate

In the skip-connection path, we propose to add multi-dilated attention gate to HF-UNet. Multi-dilated attention gate are added to the skip-connection path between the encoder and the decoder, which helps to suppress unimportant features during the encoding process and enhance valuable features. The multi-dilated attention gate is shown in Fig. 5.

Multi-dilated attention gate uses dilation convolution for extraction of global and local information. To obtain global features, we use dilation convolution with dilations of 5 and 7 for obtaining global information. To obtain local features, we use dilation convolution with dilation of 1 and 2 for obtaining local information. The extracted global and local features are batch normalized and then concat fused to obtain a number of channels that is 4 times the original number of channels. After a Relu activation layer, the input is fed to the voting module (VM),

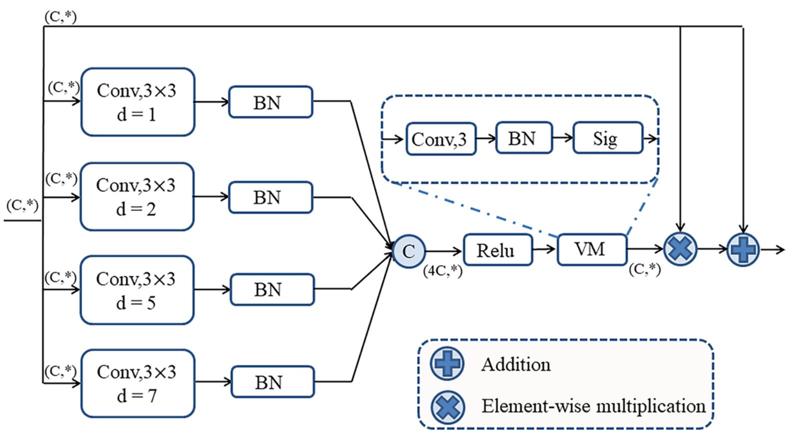

Figure 5. Components of the multi-dilated attention gate.

which filters out the valuable feature information and the number of channels is restored to the original number of channels. Finally, the elemental multiplication and summation with the original input is output to the decoder. The voting module (VM) consists of a 3×3 standard convolution, batch normalization and Sigmoid activation function. The above process can be expressed in the following equation:

$$x_1, x_2, x_3, x_4 = D_1Conv(x), D_2Conv(x), D_5Conv(x), D_7Conv(x), \tag{13}$$

$$X = Relu\{Concat[BN(x_1), BN(x_2), BN(x_3), BN(x_4)]\}, \tag{14}$$

$$V_x = Sig\{BN[Conv(X)]\}, \tag{15}$$

$$Out = x + x \cdot V_x, \tag{16}$$

where D_1Conv, D_2Conv, D_5Conv, D_7Conv denote 3×3 dilated convolutions with dilations of 1, 2, 5, and 7, respectively, x is the input, BN denotes the batch normalization, $Concat$ denotes the cascade operation, $Conv$ denotes the 3×3 standard convolution, and $Relu$ and Sig denote the Relu activation function and the Sigmoid activation function, respectively.

Experiments
Implementation details
The segmentation experiments for oral ulcers were all performed on our proposed Autooral dataset. We randomly assigned the Autooral dataset into training set (70% of the total), validation set (10% of the total) and test set (20% of the total) by patient. The experiments were all implemented based on Python 3.8 and Pytorch 1.12.0. Our experiments are implemented on a single NVIDIA GeForce RTX 4080 Laptop GPU with 12 GB of memory. For training data, we used data enhancement operations[14,19,20,32] (horizontal flip, vertical flip and random rotation) for improving data diversity. The image size is uniformly 256×256, the training epoch is set to 250, and the batch size is 8. The loss function uses the BceDice loss function[20]. The optimizer uses AdamW[33], the initial learning rate size is set to 0.001, the minimum learning rate is set to 0.00001, and the cosine annealing learning rate scheduler is used.

Evaluation metrics
In the field of medical image segmentation, several commonly used evaluation criteria are mean dice similarity coefficient (DSC), accuracy (ACC), sensitivity (SE), and specificity (SP). DSC is mainly used to measure the similarity between predicted masks and ground truth. ACC is mainly used to measure the percentage of correct classifications. SE is used to measure the percentage of true positives (TP) among true positives (TP) and false negatives (FN). SP is used to measure the percentage of true negatives (TN) in true negatives (TN) and false positives (FP).

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{17}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{18}$$

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{19}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{20}$$

where TP denotes true positive, FP denotes false positive, FN denotes false negative and TN denotes true negative.

Comparison with SOTA methods
In order to demonstrate the effectiveness of our model, we compare our experimental results with 12 of the most popular medical image segmentation models. They are UNet[9], Att U-net[18], SCR-Net[34], TransNorm[19], MALUNet[20], C^2SDG[24], M^2SNet[22], MSA[26], META-Unet[35], MHorUNet[14], VM-UNet[36] and H-vmunet[32].

As Table 2 shows our experimental results. In the table we can see that our proposed HF-UNet has the best performance. The DSC value of TransNorm is the lowest because TransNorm is based on the model of Transformers, which has the disadvantage of requiring large training samples. Our HF-UNet gets a DSC value of almost 0.80 on the Autooral dataset. As can be seen from the visualization of the segmentation results shown in Fig. 6, our model achieves state-of-the-art performance both for large ulcers and for small and densely distributed ulcers. This is an example of an ulcer with a fuzzy boundary and strong interference from teeth and reflections, as shown in Fig. 6F. The prediction of the proposed HF-UNet is the closest result to the doctor's labeling, and all other models are misled by the blurred boundary and other interference terms. In particular, the upper part of the ulcer gets a correct contour orientation only by our proposed HF-UNet.

In addition, we also give the computational complexity (GFLOPs) and inference memory usage of the model in Table 2. From the table, it can be concluded that although the GFLOPs of our proposed HF-UNet are slightly

Methods	Publication year	GFLOPs [↓]	Merrmory usage (MB) [↓]	DSC [↑]	ACC [↑]	SP [↑]	SE [↑]
UNet[9]	2015	3.224	1567	0.7480	0.9617	0.9815	0.7282
Att U-net[18]	2018	8.575	1580	0.7404	0.9632	0.9879	0.6716
SCR-Net[34]	2021	1.567	1569	0.7069	0.9602	0.9896	0.6148
TransNorm[19]	2022	39.284	2113	0.5670	0.9514	0.9873	0.4691
MALUNet[20]	2022	**0.083**	1551	0.6318	0.9409	0.9655	0.6500
C^2SDG[24]	2023	7.972	1723	0.7210	0.9604	0.9862	0.6554
M^2SNet[22]	2023	9.026	1753	0.7482	0.9669	**0.9953**	0.6308
MSA[26]	2023	402.893	5173	0.7540	0.9697	0.9887	0.7181
META-Unet[35]	2023	5.139	1639	0.7535	0.9626	0.9822	0.7318
MHorUNet[14]	2024	0.864	1597	0.7618	0.9657	0.9867	0.7180
VM-UNet[36]	2024	4.112	1124	0.7639	0.9636	0.9811	**0.7555**
H-vmunet[32]	2024	0.742	**1060**	0.7127	0.9605	0.9887	0.6276
HF-UNet (Our)	2024	10.715	2029	**0.7972**	**0.9715**	0.9932	0.7257

Table 2. Performance comparison on Autooral dataset. Significant values are in bold.

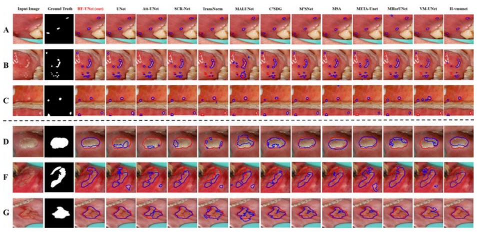

Figure 6. Visual segmentation results in the Autooral dataset. The red contour line indicates the ground truth and the blue contour line indicates the model prediction split line. Regarding the classification of ulcer size, cases (**A**), (**B**) and (**C**) are characterized by ulcers that are individually tiny in size and widely distributed, whereas cases (**D**), (**E**) and (**F**) have larger ulcers. Regarding the classification of ulcer location, cases (**A**), (**B**), (**D**) are the result of lesion segmentation in the tongue area and cases (**C**), (**F**) and (**G**) are the result of lesion segmentation in the non-tongue area.

larger, our proposed HF-UNet allows reasoning at a single graphics card memory of around 2048 MB and yields the most excellent oral ulcer segmentation results. In particular, MSA has the highest GFLOPs and inference memory usage due to its use of the segment anything model (SAM)[25] as its underlying framework.

In addition, we visualized the lesion segmentation results for different ulcer size cases. As shown in Fig. 6, the ulcers in A, B and C are characterized by tiny and widely distributed individual areas, while the ulcer areas in D, E and F are much larger. By comparing the visualization graphs of the segmented tiny and widely distributed ulcers of each model, it can be concluded that the proposed HF-UNet is able to identify the widely distributed tiny ulcers better. The other compared models have problems such as recognition omission and boundary prediction bias in recognizing tiny and widely distributed ulcers. Further, oral ulcers generally appear in the tongue region as well as in the non-tongue region (areas such as lips and soft palate). We visualized the lesion segmentation results of different models according to the tongue area as well as the non-tongue area. As shown in Fig. 6A,B,D are the lesion segmentation results in the tongue region, and C, F and G are the lesion segmentation results in the non-tongue region. In particular, in the example of D in the tongue region, our proposed HF-UNet has more complete and accurate boundary prediction results. In addition, in the C example of the non-tongue area, the remaining comparison models suffer from recognition omissions, while the proposed HF-UNet is able to better recognize the tiny and widely distributed ulcers in both the tongue and non-tongue areas.

Ablation study

To further validate the effectiveness of our proposed high-order focus interaction module (HFblock) and a lesion localization module (LL-M), we performed a series of ablation experiments. The validation was carried out by using LL-M only, HFblock only, LL-M+HFblock input data flow method and HFblock+LL-M input data flow method. Figure 7a shows the results of our experiments, from which we can see that the best performance is achieved by using the input data flow approach of HFblock+LL-M, with a DSC value of nearly 0.80. The use of only one module alone leads to a degradation of the model performance.

In addition, we set up several experiments in order to verify the performance impact of different orders of focus interaction modules. As shown in Fig. 7b, e.g. [2, 3, 4] indicates that the encoder sets the HFblock module for 2, 3, and 4 order focus interactions according to the direction of the input data flow, respectively. The HFblock

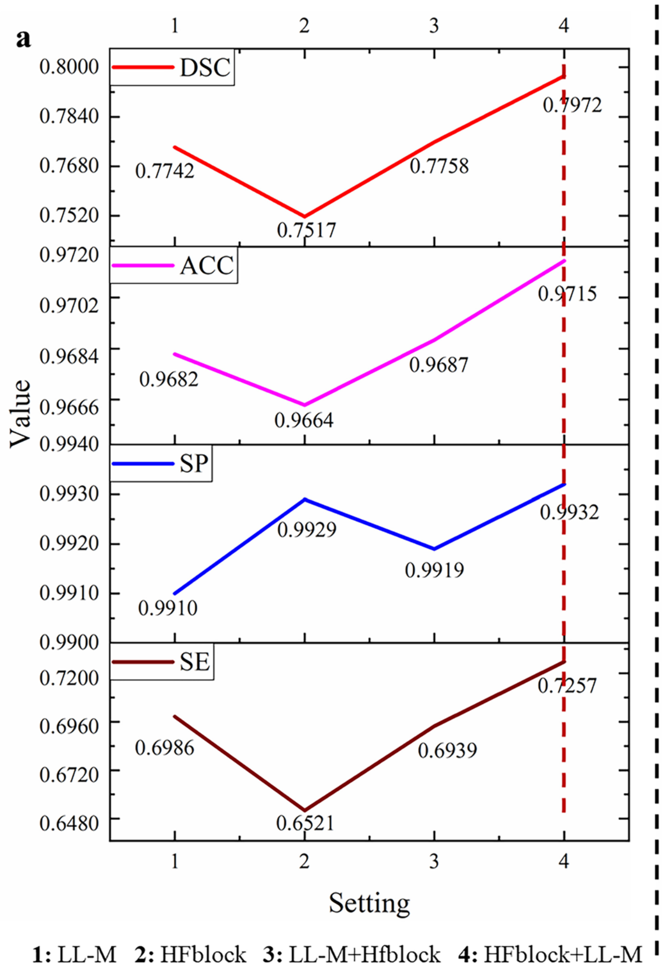

Figure 7. (**a**) Ablation experiments on the effectiveness of different combinations of LL-M and HFblock in HF-UNet. (**b**) Ablation experiments on the effectiveness of HF-UNet with different order selections.

in the decoder section is kept symmetric with the encoder. Figure 7b shows our experimental results, and we can get that the setting of [4, 4, 4] achieves the best performance. In order to more intuitively observe the feature extraction at various settings, we performed visualizations. As shown in Fig. 8 is a visualization of each order of the last HFblock of the decoder when set to [2, 2, 2], [3, 3, 3], [4, 4, 4] and [5, 5, 5]. We can obtain from Fig. 8 that oral ulcers are most clearly characterized at the 4-order of [4, 4, 4]. The performance shows degradation results for the [5, 5, 5] setting. Similarly, the [3, 4, 5] setup also shows lower performance than the [2, 3, 4] setup. They all showed performance degradation at the 5-order of use. In the HF-UNet model of this study, we used the [4, 4, 4] setting. This is analyzed as follows, when a setting of 5-order of [5, 5, 5] is used, this results in too sparse a number of channels for the first input to the Focus module. Specifically, the number of first input Focus module channels at 5-order is only C/16 from Eq. (9). This results in subsequent higher order interactions that fail to perform ulcer feature learning well. More directly, as shown in Fig. 8, it is difficult to visualize the ulcer contour at 5-order, and instead, many bright spots appear on many tooth features. This indicates that the ulcer features failed to be learned well, so more number of channels should be provided initially for input to the Focus module for ulcer feature learning. Therefore, the number of channels for the first input feature focus learning should not be less than C/16. In the experiments of this study, we chose the [4, 4, 4] setting for our model.

Discussion and limitations

Mouth ulcers differ from other diseases in that they have strong interference terms, which include similar dental features and teeth that provide more reflexive information (Fig. 1). Therefore, we aim to propose a high-quality annotated oral ulcer dataset (Autooral), where high-quality 420 images can already train the model well (DSC of 0.80). The full age coverage, 13-year collection interval, and the presence of 12 underlying diseases, among others, show that we have enough sample diversity to provide model training data support. Our original intention of proposing multitasking was to provide more tasks for the dataset, which provides future researchers with richer studies. However, our technical model (HF-UNet) is mainly focused on segmentation, because high quality labeling for segmentation tasks is our original intention, and classification tasks are complementary.

In order to better cope with the various disturbances of oral ulcers, we propose HF-UNet for oral ulcer segmentation. We find better resistance to this interference information in the frequency domain (GLF), which

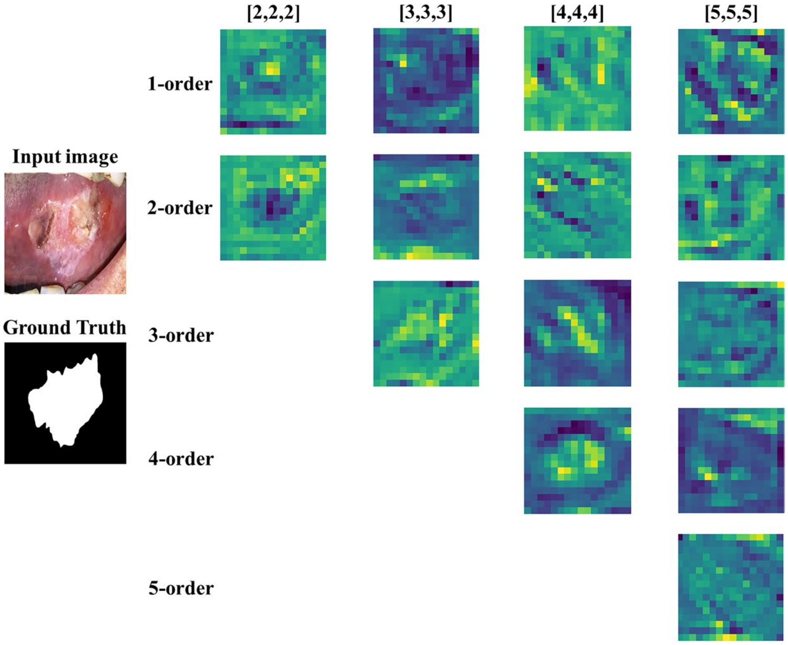

Figure 8. Result graphs for each order of the last HFblock of the decoder set to [2, 2, 2], [3, 3, 3], [4, 4, 4] and [5, 5, 5].

also coincides with Rao et al.[29]. However, it is difficult to extract detailed information on ulcers by relying solely on GLF. Therefore, we used the Focus module for feature extraction of local details. However, if all channels of the global information extracted by GLF were immediately fused with the local information extracted by the Focus module, excessive redundancy would be introduced. So this is why it is important to introduce higher-order interactions. Higher-order interactions can introduce global information incrementally while extracting local information at each step. This minimizes the input of redundant information in the global information. As shown in Figs. 7b and 8, the best performance is achieved when 4-order is performed, with the least amount of redundant information. Surprisingly, the MHorUNet (similar to the high-order form) proposed by the previous authors in Table 2 has the second highest performance, which reaffirms the validity of our proposed high-order focus interaction block (HFblock).

Although our proposed multi-task oral ulcer dataset (Autooral) and HF-UNet are exciting, in this study, we have only performed a detailed study on the segmentation task of the proposed Autooral dataset. In the future, an in-depth study of the classification task in the Autooral dataset or even a detailed study of oral ulcers combining the classification and segmentation tasks is an important aspect. In addition, the incorporation of the proposed Autooral dataset with more clinical cases under different geographical regions to enhance its comprehensiveness and generalization to a wider population is also an important direction. Furthermore, applying different learning strategies to the proposed HF-UNet and exploring it in more medical image segmentation tasks is also an important direction.

Conclusions

In this paper, we present a high-quality multi-tasking oral ulcer dataset (Autooral) containing segmentation and classification tasks. We make it publicly available for bridging the gap of public oral ulcer datasets in the research field. To the best of our knowledge, we are the first team to make publicly available a multi-tasking oral ulcer dataset. We also propose a novel model architecture HF-UNet for oral ulcer segmentation. The proposed high-order focus interaction module (HFblock) combines the acquisition global property of high-order attention with the acquisition local property of focused interaction. We visualize each order inside the HFblock, and lesion features gradually become clear, prominent and complete. In addition, the proposed LL-M can enhance the ability of the model to detect the edges of lesions. Experimental results show that the proposed dataset (Autooral) in the proposed HF-UNet achieves a DSC of nearly 0.80, while the current state-of-the-art model only has a DSC value of around 0.76. We hope that our proposed work will facilitate further research in computer vision in oral mucosal medicine.

Data availability

The proposed Autooral dataset link and the HF-UNet model code are available from https://github.com/wurenkai/HF-UNet-and-Autooral-dataset.

Received: 8 June 2024; Accepted: 31 July 2024
Published online: 29 August 2024

References

1. Zeng, X. *et al.* Difficult and complicated oral ulceration: An expert consensus guideline for diagnosis. *Int. J. Oral Sci.* **14**, 28 (2022).
2. Guo, G. & Razmjooy, N. A new interval differential equation for edge detection and determining breast cancer regions in mammography images. *Syst. Sci. Control Eng.* **7**, 346–356 (2019).
3. Liu, Q., Liu, Z., Yong, S., Jia, K. & Razmjooy, N. Computer-aided breast cancer diagnosis based on image segmentation and interval analysis. *Automatika* **61**, 496–506 (2020).
4. Minhas, S. *et al.* Oral ulcers presentation in systemic diseases: An update. *Open Access Maced. J. Med. Sci.* **7**, 3341 (2019).
5. dos Santos, F. D. S. *et al.* Misdiagnosis of lip squamous cell carcinoma. *RSBO Rev. Sul-Bras. Odontol.* **9**, 114–118 (2012).
6. Mortazavi, H. *et al.* Diagnostic features of common oral ulcerative lesions: An updated decision tree. *Int. J. Dent.* **2016**, 7278925 (2016).
7. Valente, V. B. *et al.* Oral squamous cell carcinoma misdiagnosed as a denture-related traumatic ulcer: A clinical report. *J. Prosthet. Dent.* **115**, 259–262 (2016).
8. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proc. of the IEEE conference on computer vision and pattern recognition*, 3431–3440 (2015).
9. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241 (Springer, 2015).
10. Yang, J. *et al.* Focal self-attention for local-global interactions in vision transformers. Preprint at http://arxiv.org/abs/2107.00641 (2021).
11. Naderi, M., Givkashi, M., Piri, F., Karimi, N. & Samavi, S. Focal-unet: Unet-like focal modulation for medical image segmentation. Preprint at http://arxiv.org/abs/2212.09263 (2022).
12. Yang, J., Li, C., Dai, X. & Gao, J. Focal modulation networks. *Adv. Neural Inf. Process. Syst.* **35**, 4203–4217 (2022).
13. Rao, Y. *et al.* Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Adv. Neural Inf. Process. Syst.* **35**, 10353–10366 (2022).
14. Wu, R. *et al.* Mhorunet: High-order spatial interaction unet for skin lesion segmentation. *Biomed. Signal Process. Control* **88**, 105517 (2024).
15. Anantharaman, R., Velazquez, M. & Lee, Y. Utilizing mask r-cnn for detection and segmentation of oral diseases. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2197–2204 (IEEE, 2018).
16. Jain, M., Rai, C. *et al.* Early detection of oral ulcers using photographic evidence: A novel approach using ensemble of convolution neural network. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, 1–5 (IEEE, 2022).
17. Ding, H., Huang, Q. & Rodriguez, D. Modified locust swarm optimizer for oral cancer diagnosis. *Biomed. Signal Process. Control* **83**, 104645 (2023).
18. Oktay, O. *et al.* Attention u-net: Learning where to look for the pancreas. Preprint at http://arxiv.org/abs/1804.03999 (2018).
19. Azad, R., Al-Antary, M. T., Heidari, M. & Merhof, D. Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. *IEEE Access* **10**, 108205–108215 (2022).
20. Ruan, J., Xiang, S., Xie, M., Liu, T. & Fu, Y. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1150–1156 (IEEE, 2022).
21. Ullah, Z., Usman, M., Jeon, M. & Gwak, J. Cascade multiscale residual attention cnns with adaptive ROI for automatic brain tumor segmentation. *Inf. Sci.* **608**, 1541–1556 (2022).
22. Zhao, X. *et al.* M^2snet: Multi-scale in multi-scale subtraction network for medical image segmentation. Preprint at http://arxiv.org/abs/2303.10894 (2023).
23. Ullah, Z., Usman, M., Latif, S. & Gwak, J. Densely attention mechanism based network for covid-19 detection in chest x-rays. *Sci. Rep.* **13**, 261 (2023).
24. Hu, S., Liao, Z. & Xia, Y. Devil is in channels: Contrastive single domain generalization for medical image segmentation. Preprint at http://arxiv.org/abs/2306.05254 (2023).
25. Kirillov, A. *et al.* Segment anything. Preprint at http://arxiv.org/abs/2304.02643 (2023).
26. Wu, J. *et al.* Medical sam adapter: Adapting segment anything model for medical image segmentation. Preprint at http://arxiv.org/abs/2304.12620 (2023).
27. Chu, X. *et al.* Twins: Revisiting the design of spatial attention in vision transformers. *Adv. Neural Inf. Process. Syst.* **34**, 9355–9366 (2021).
28. Wang, S., Li, B. Z., Khabsa, M., Fang, H. & Ma, H. Linformer: Self-attention with linear complexity. Preprint at http://arxiv.org/abs/2006.04768 (2020).
29. Rao, Y., Zhao, W., Zhu, Z., Lu, J. & Zhou, J. Global filter networks for image classification. *Adv. Neural Inf. Process. Syst.* **34**, 980–993 (2021).
30. Ning, Z., Zhong, S., Feng, Q., Chen, W. & Zhang, Y. Smu-net: Saliency-guided morphology-aware u-net for breast lesion segmentation in ultrasound image. *IEEE Trans. Med. Imaging* **41**, 476–490 (2021).
31. Lin, Y. *et al.* Rethinking boundary detection in deep learning models for medical image segmentation. In *International Conference on Information Processing in Medical Imaging*, 730–742 (Springer, 2023).
32. Wu, R., Liu, Y., Liang, P. & Chang, Q. H-vmunet: High-order vision mamba unet for medical image segmentation. Preprint at http://arxiv.org/abs/2403.13642 (2024).
33. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at http://arxiv.org/abs/1711.05101 (2017).
34. Wu, H., Zhong, J., Wang, W., Wen, Z. & Qin, J. Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos. *Proc. AAAI Conf. Artif. Intell.* **35**, 2916–2924 (2021).
35. Wu, H., Zhao, Z. & Wang, Z. Meta-unet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation. *IEEE Trans. Autom. Sci. Eng.* https://doi.org/10.1109/TASE.2023.3292373 *(2023).*
36. Ruan, J. & Xiang, S. Vm-unet: Vision mamba unet for medical image segmentation. Preprint at http://arxiv.org/abs/2402.02491 (2024).

Acknowledgements

This project was supported by the National Natural Science Foundation of China (Nos. 81970941), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD, 2018-87), and the Critical Research and Development Project of Jiangsu Province (Nos. BE2021723).

Author contributions

C.J. provided and analyzed data, assisted in designing experiments and wrote the first draft. R.W. designed the algorithms, experiments and wrote the first draft. Y.L. and Y.W. assisted in analyzing the data and plotting the graphs. Q.C. provided the conditions for the experiments as well as financial support. P.L. planned the overall project and assisted in the design of the algorithms. Y.F. planned the overall project and provided the data support as well as the financial support. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to P.L. or Y.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.