# *MSDEnet: Multi-scale detail enhanced network based on human visual system for medical image segmentation*

Computers in Biology and Medicine 170 (2024) 108010

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# MSDEnet: Multi-scale detail enhanced network based on human visual system for medical image segmentation

Yuangang Ma [a], Hong Xu [a,b,*], Yue Feng [a], Zhuosheng Lin [a], Fufeng Li [c], Xin Wu [a], Qichao Liu [a], Shuangsheng Zhang [d]

[a] *Department of Intelligent Manufacturing, Wuyi University, China*
[b] *Victoria University, Australia*
[c] *Syndrome Laboratory of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, China*
[d] *Jiangmen Central Hospital, China*

ARTICLE INFO

*Index terms:*
Medical image segmentation
Convolutional neural network (CNN)
Multi-scale structure
Redundant information

ABSTRACT

In medical image segmentation, accuracy is commonly high for tasks involving clear boundary partitioning features, as seen in the segmentation of X-ray images. However, for objects with less obvious boundary partitioning features, such as skin regions with similar color textures or CT images of adjacent organs with similar Hounsfield value ranges, segmentation accuracy significantly decreases. Inspired by the human visual system, we proposed the multi-scale detail enhanced network. Firstly, we designed a detail enhanced module to enhance the contrast between central and peripheral receptive field information using the superposition of two asymmetric convolutions in different directions and a standard convolution. Then, we expanded the scale of the module into a multi-scale detail enhanced module. The difference between central and peripheral information at different scales makes the network more sensitive to changes in details, resulting in more accurate segmentation. In order to reduce the impact of redundant information on segmentation results and increase the effective receptive field, we proposed the channel multi-scale module, adapted from the Res2net module. This creates independent parallel multi-scale branches within a single residual structure, increasing the utilization of redundant information and the effective receptive field at the channel level. We conducted experiments on four different datasets, and our method outperformed the common medical image segmentation algorithms currently being used. Additionally, we carried out detailed ablation experiments to confirm the effectiveness of each module.

## 1. Introduction

Accurately locating lesions or abnormal information in a large number of medical images is a challenging task for physicians [1]. Accurate medical image segmentation models can significantly reduce workload and pressure for physicians [2]. In recent years, various types of medical image segmentation models have emerged, among which Unet [3] is widely used due to its excellent multi-layer information fusion capability. Different feature layers possess different feature information, deeper feature maps contain richer semantic information and focus more on the location and shape of the target object, while shallow feature maps contain more detailed information, facilitating accurate boundary determination [4]. Unet fuses feature information from different layers using skip connections to generate precise prediction results [3]. Combined with various modules that enhance the feature extraction capability of the network, such as transformer [5,6], MLP [7], and attention mechanism [8,9], various types of medical image segmentation models with Unet-based architecture have been frequently proposed.

However, conventional medical image segmentation models have two limitations. Firstly, as shown in Fig. 1, there are high redundancy and similarity in feature information across channels. The channel dimension of the feature map continuously expands throughout the encoder stage, leading to an increase in redundant information. Frequent cross-level feature fusion operations in Unet-shape also amplify the impact of redundancy, diluting important details of information and affecting the ability to detect the precise location of the target object. The multi-scale subtraction unit (MSU) is designed to reduce the impact of redundant information and obtain rich multi-scale difference information [4]. Nevertheless, this approach does not explicitly address the issue of reducing redundant information at the source. Secondly, single-scale convolutional kernels limit the network's ability to obtain

[*] Corresponding author at: Victoria University, Australia.
*E-mail addresses:* hong.xu@vu.edu.au (H. Xu), J002443@wyu.edu.cn (Y. Feng).

https://doi.org/10.1016/j.compbiomed.2024.108010
Received 17 August 2023; Received in revised form 24 December 2023; Accepted 15 January 2024
Available online 18 January 2024

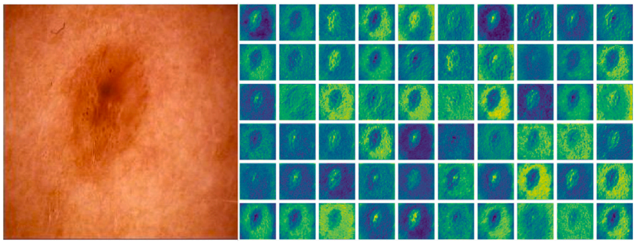


**Fig. 1.** Redundant information across channels.

target contextual information [10]. Based on the scale-space theory, the superposition of convolutional kernels of multiple sizes can collect subtle size variations and detailed information, bringing better feature extraction capability to the network. Therefore, some studies [11–14] have designed parallel multi-branch networks to introduce multi-scale information to the network, while others [3,15] have used serial skip connection networks to accomplish the fusion of multi-scale information data. However, these methods generally use standard convolution within a single scale, which may limit the feature extraction capability of multi-scale structures.

To address the aforementioned issues, a new multi-scale detail enhanced network (MSDEnet) was proposed. First, for redundant information, the channel multi-scale (CMS) module was improved from Res2net module. CMS employs independent multi-scale branch convolution operations for feature extraction to reduce the impact of redundant feature information. To enhance the ability of the network to extract multi-scale details, the detail enhanced (DE) module was designed, inspired by human visual features. The peripheral information can help determine the relative position of the observation target, and can also be contrasted with the central information to highlight features such as the detailed information of the observation target. In this study, two different directions of asymmetric convolution and a standard convolution superposition are used to enhance the skeleton of the convolution kernel. The structure is characterized by a strong central and weak peripheral distribution within the convolutional kernel, which is similar to the distribution of the central and peripheral receptive fields in humans. This characteristic is beneficial to the feature extraction ability of the multi-scale structure. Based on this, we extend the structure to the multi-scale level, and the multi-scale detail enhanced (MSDE) module was constructed. By collecting information at different scales, better segmentation performance can be further achieved.

The main contributions of this study are summarized as follows:

(1) We present the Multi-scale detail enhanced network (MSDEnet). The MSDEnet benefits from a unique coding structure and detail enhancement mechanism, which improves the network's sensitivity to detailed information and achieves higher precision medical image segmentation. Our experiments across four datasets showcase optimal performance, with achievements on 15 out of 20 metrics.
(2) We present the Multi-scale detail enhanced module (MSDE). The MSDE module contains multiple DE modules at different scales, and the structure creates diffuse comparison of center and periphery information, which can capture small feature information changes more acutely.
(3) We present the Channel multi-scale module (CMS). The CMS module reduces the generation of redundant information from the encoder stage by means of an independent and parallel multi-scale structure within the channel. This structure improves the utilization efficiency of redundant information while reducing the number of times of feature information reuse, avoiding the redundant information in the subsequent addition or concatenation, the important information of the boundary is seriously diluted, thus leading to boundary blurring.

## 2. Related work

### 2.1. Medical image segmentation network

In the field of medical image segmentation, Unet [3] is proposed to solve the problem that medical images usually contain noise and have blurred boundaries. The structure proves to be the most efficient and versatile medical image segmentation backbone network available. Since then, there has been a proliferation of improved and complementary network structures based on the Unet. Unet++ [16] uses dense connections to link all layers in the network together, and the decoder feature information is progressively enriched before being passed into the encoder, which can effectively capture the fine-grained details of foreground objects.

The attention mechanism draws on the characteristics of human vision to assign more weight to more important information. Thereafter, researchers have continued to propose multiple network models based on Unet that incorporate various types of attention mechanisms [8,17–19]. For example, Att-Unet [8] incorporates an attention module to filter the feature information from the encoder before the decoder fuses the shallow feature information, suppressing irrelevant regions in the input image while highlighting salient features in specific local regions to improve the sensitivity of the model and the accuracy of prediction.

After that, the Transformer [20] changes the situation of CNN domination in the field of computer vision. Because of the receptive field limitation, CNN cannot make good use of global information but has excellent local information extraction ability, while the Transformer is the opposite of CNN. TransU-net [5] combines both, using CNN for shallow feature extraction, and then converting the feature map into token for global information encoding. After transformer, MLP-mixer [21] is introduced to bring fully connected structures into the field of computer vision once again, with lower inductive bias. Unext [7] introduces and re-engineers the MLP to maintain performance while reducing the number of parameters.

Simultaneously, the novel research based on the fusion of CNNs, transformer and MLP as, combined with cloud computing, image statistical feature information and other techniques, provides reference and thinking for the field [22–25].

### 2.2. Multi-scale structure

In the real world, many objects and structures have multi-scale characteristics, meaning they exhibit different properties or features at different scales or levels of detail [26]. To better capture these multi-scale characteristics in computer vision models, scale space theory has been proposed and employed. This theory suggests that incorporating multi-scale convolution kernels into models can help better capture the key characteristics of images at multiple scales [27]. There are many types of multi-scale structures that have been developed based on this theory. These structures can be divided into intra-layer multi-scale structures and inter-layer multi-scale structures based on their form. Intra-layer multi-scale structures are usually in the form of plug-in modules that can be integrated into existing models to allow them to extract features at multiple scales, such as Inception [12], ASPP [28], DenseASPP [29] and RFB [30]. The latter is mainly reflected in the codec network for end-to-end feature fusion, such as the Unet family [3,5,7,10,16,31,32].

Res2net [26] introduces a multi-scale structure to the channel level, which is different from previous multi-scale structures and does not increase the computational cost significantly. However, the multi-scale branches in the Res2net module are not independent of each other, which can lead to redundant information reuse in different branches. To address this limitation, CMS was proposed in this study. CMS separates the different multi-scale branches and increases the size of the convolution kernel.

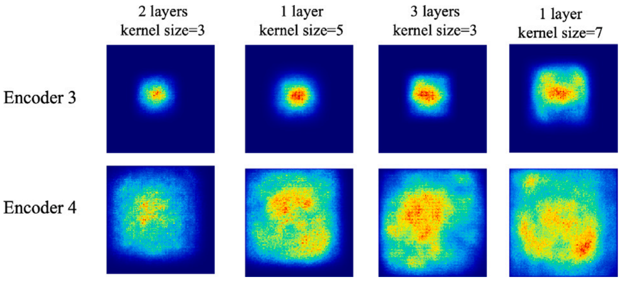


**Fig. 2.** Comparing the effect about number of layers and size of kernel.

### 2.3. Effective receptive field

For segmentation tasks, the receptive field is a concept of great interest. Just like the human visual field, the receptive field reflects the perceived range of the convolutional kernel for the current input feature map. If the range is small, then the information received is one-sided and localized. If the receptive field is increased, then more global information can be obtained, which is more conducive to the judgment of the current situation [15,33–37].

Most existing CNN models expand the size of the receptive field by stacking convolutional kernels with pooling layers. However, the size of the effective receptive field (ERF) is related to the size of the convolutional kernel $K$ and the depth of the model $L$, which is proportional to $K$ and inversely proportional to $L$ [38]. In Fig. 2, the ERF is more sensitive to the size of the convolutional kernel, and increasing the depth is not as intuitive as increasing the size of the convolutional kernel. In addition, increasing the depth causes optimization issues. Although the residual mechanism solves the limitation of network degradation, the ERF of stacking small convolutional kernels is still not necessarily large [39]. Thus, this study replaced the superposition of 3 × 3 convolution in the Res2net module with a large convolution kernel to obtain a larger ERF. To reduce parameters and a floating point operations per second (FLOPS) associated with large convolution kernels, the deeply separable convolution was used in this study.

## 3. Method

As shown Fig. 3, the MSDEnet architecture includes five CMS as encoder for multi-level feature extraction. The output feature information is then input into MSDE. In the MSDE, the feature information is first passed through $n$ ($n$ is the number of multi-scale scales) independent and different scale DE. The results are concatenated and passed to the decoder.

The decoder gradually performs end-to-end feature fusion from bottom-up, and the bilinear interpolation method is selected for upsampling throughout the procedure. The output of each layer at the decoding end is supervised during training. In Fig. 3, the Decoder consists of two parts: (i) bilinear interpolation of the output characteristic maps of the lower layers, which are then concatenated with the output characteristic maps of the jump connections of this layer; (ii) two consecutive 3 × 3 convolutions.

Due to the characteristics of medical image segmentation tasks, there are usually large differences between different datasets, such as data distribution or background differences. In order to adapt to the those, we choose to use Bce-Dice Loss functions. Because Bce Loss has better stability at the beginning of training and helps to speed up the convergence of the model. Dice Loss, on the other hand, focuses more on pixel-level similarity, which can motivate the model to generate smoother and continuous segmentation results, and Bce-Dice loss function combines the advantages of both. This function is listed as follows:

$$J(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + \left(1-y_n\right)\log\left(1-\hat{y}_n\right)\right] \tag{1}$$

The loss of prediction maps and real labels in each layer is used as the total loss, and weights are assigned to the prediction results of different depths according to the ratio of 1 : 0.5 : 0.25 : 0.125, and the deeper the depth, the smaller the weighting coefficients are.

Finally, the output results of the final 4-layer decoder are jointly passed into the loss function for supervision.

### 3.1. Channel multi-scale module

In medical image segmentation network, there are frequent additions or concatenations, which may amplify the impact of the redundant information generated during feature extraction. Excessive redundant information can dilute important boundary information and lead to blurred boundaries in segmentation results.

The channel multi-scale module structure is shown in Fig. 4(b). We make targeted improvements to the Res2net module with the aim of optimizing the reuse of redundant information and increasing ERF in it. As in the Res2net module, we divide the input feature map into input subsets, denoted as $x_i$, where $i \in \{1, 2, \dots, s\}$. The difference is that no information fusion between subsets is performed in the CMS. Each $x_i$ is convolved by $Conv_{(2i-1)\times(2i-1)}$ to obtain the corresponding output subset $y_i$. The process can be expressed as:

$$y_i = Conv_{(2i-1)\times(2i-1)}\left(x_i\right) \quad i = 1, 2, 3, \dots, s. \tag{2}$$

The advantage of CMS is that it not only reduces the generation of redundant information, but also makes full use of the existing redundant information for multi-scale feature extraction as much as possible. The cascaded 3 × 3 convolution is replaced with a larger convolution kernel that is more favorable for segmentation, which gives the backbone a larger ERF. The introduction of depthwise separable convolution, which consists of depthwise convolution (DW) and pointwise convolution (PW), reduces the increase in computational cost associated with large convolution kernels while ensuring accuracy.

### 3.2. Multi-scale detail enhanced module

People have two types of vision, central vision and peripheral vision. Central vision is used to look directly at things to observe details, while peripheral vision shows other areas of the visual field, that is, the peripheral area that the human eye can see [40]. Brain imaging studies have found that input from the peripheral visual field can be decoded in the primary visual cortex, which characterizes the central visual field, suggesting the existence of a top-down peripheral-central feedback mechanism in visual discrimination tasks such as shape, color, and object category judgments [41]. The study found that the use of asymmetric convolution in different directions and standard convolution superimposed in parallel can produce different degrees of attention inside and outside on the receptive field, which is very similar to human central and peripheral vision. As a result, the detail enhanced module is designed.

Usually in most studies, the superposition of $(2i-1)\times 1$ and $1\times(2i-1)$ asymmetric convolution replace standard convolution in order to improve the inference speed without reducing the representational power, as in Inception-V3 [42]. On the other hand, the use of 3 × 1, 1 × 3, and 3 × 3 superimposed to enhance the skeleton of convolution kernel to improve feature extraction has been shown to be effective compared to 3 × 3 convolution [39]. We then extend it to $(2i-1)\times[2(i-1)-1]$ and $[2(i-1)-1]\times(2i-1)$ as the skeleton reinforcement of the large convolution kernel. For several 2D convolutional kernels acting on the same input, with the same stride and compatible size, we can add the parameters of these kernels at the corresponding positions to obtain an equivalent convolutional kernel to obtain the same output. This is the additivity of 2D convolution kernels [43],

$$I * K_1 + I * K_2 + \cdots + I * K_n = I * \bigoplus_{i=1}^{n} K_i \tag{3}$$

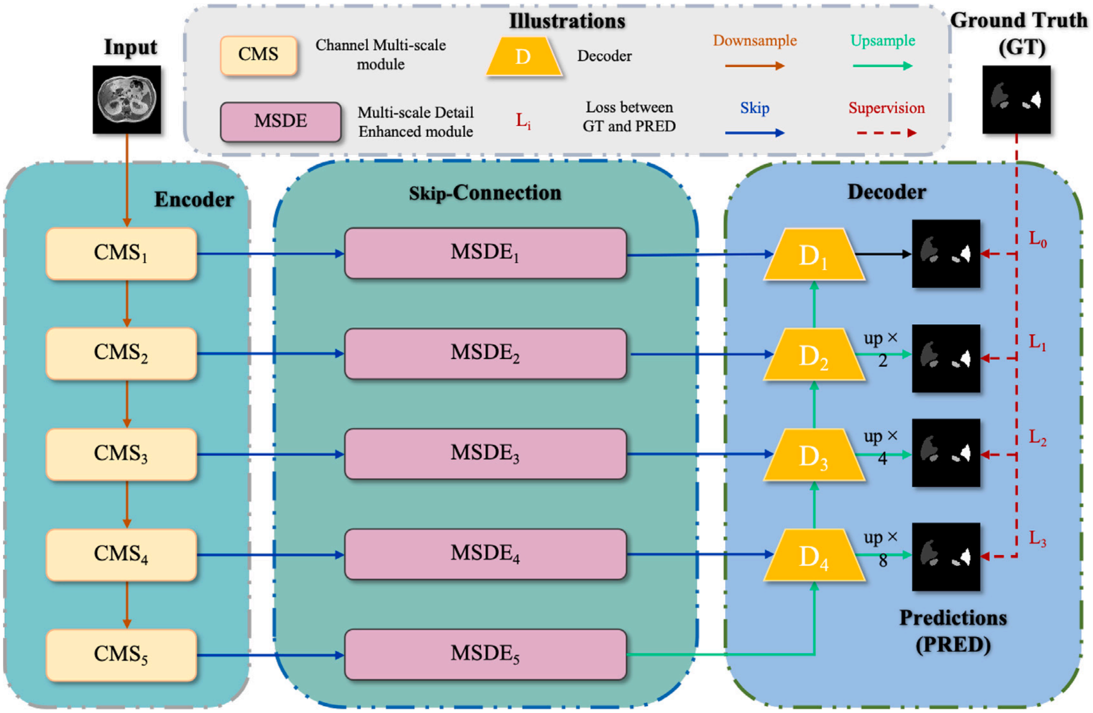
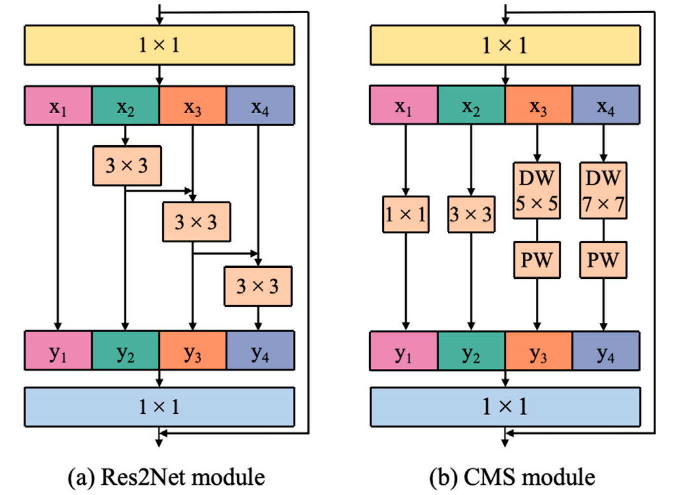


**Fig. 3.** Overview of the MSDEnet architecture.



(a) Res2Net module (b) CMS module

**Fig. 4.** Comparison between the Res2net module and CMS module.

where $I$ is a input matrix, $K_i$ is compatible sizes 2D convolution kernels, $n$ is the number of $K$, and $\oplus$ is the corresponding positions element-wise addition of the kernel parameters. Here compatible means that we can "padding" or "patch" the smaller convolutional kernels into larger ones. Formally, We denote the larger convolutional kernel by $L$ and the smaller convolutional kernel by $S$, so

$$H_L \geq H_S,\ W_L \geq W_S,\ M_L = Patch\left(M_S\right) \tag{4}$$

where $H$ and $W$ represent the height and width of the convolution kernel, respectively, and $M$ represents the shape of the convolution kernel. E.g., 3 × 1 and 1 × 3 kernels are compatible with 3 × 3.

We split $DE$ into three separate and parallel branches. For $(2i-1)\times(2i-1)$ convolution branch, we call it an $EB$ (Enhanced Base). Similarly, For $(2i-1)\times(2i-1)$ and $(2i-1)\times(2i-1)$ convolution branches, we call them $E_H$ (Enhanced Horizontal) and $E_V$ (Enhanced Vertical) respectively. For $DE$, $F_{in} \in \mathbb{R}^{C\times H\times W}$ as input and we can express it as follows:

$$DE = Concat\left(E_B\left(F_{in}\right),\ E_H\left(F_{in}\right),\ E_V\left(F_{in}\right)\right) \tag{5}$$

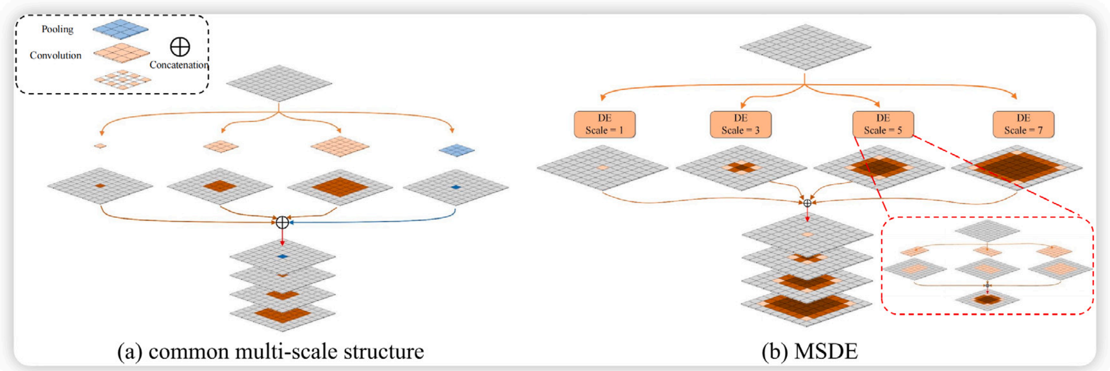
It is obvious that the output of $DE$ consists of three parts, the $CP$ (the Central Part), $PSP$ (the Peripheral Skeleton Part) and $EP$ (the Edge Part). $PSP$ contains $PSP_H$ in the horizontal direction and $PSP_V$ in the vertical direction.

$$\begin{aligned}
CP &= Concat\left(E_B\left(F_{in}\right)_{CP},\ E_H\left(F_{in}\right)_{CP},\ E_V\left(F_{in}\right)_{CP}\right),\\
PSP_H &= Concat\left(E_B\left(F_{in}\right)_{PSP},\ E_H\left(F_{in}\right)_{PSP}\right),\\
PSP_V &= Concat\left(E_B\left(F_{in}\right)_{PSP},\ E_V\left(F_{in}\right)_{PSP}\right),\\
EP &= E_B\left(F_{in}\right)_{EP}
\end{aligned} \tag{6}$$

From Eq. (6), it can be seen that the closer the edge part is, the less information enhancement is obtained. And both $PSP$ and $CP$ that are on the convolutional kernel skeleton gain different degrees of enhancement. The advantage of this is that not only the skeleton of the large convolution kernel is strengthened, but also the central enhancement part of each scale contains the entire receptive field of the previous scale. From Fig. 5 we can find a clear difference between the MSDE and the commonly used multi-scale structures. In the MSDE, this structure allows the central receptive field to show a gradual enlargement and outward contrast process, which can make the structure more sensitive to detailed information.

Typically, Unet-shape accomplishes end-to-end feature information fusion by jumping connections between the encoder and decoder, gradually replenishing the missing detail information from shallow to deep. In order to obtain better texture information for detail refinement, we insert MSDE in the skip connection. Multi-scale texture enhancement is performed on the output features of each encoder layer. The flow is shown in Fig. 5(b), where the output results of DE at different scales are concatenate, and then the number of channels is adjusted and information is exchanged between channels by a 1 × 1 convolution, which can be expressed as:

$$MSDEmodule = Conv_{1\times 1}\left\{Concat\right($$



(a) common multi-scale structure (b) MSDE

**Fig. 5.** The common multi-scale structure and MSDE.

$$\begin{aligned}&DEmodule\left(F_{in}\right)_{scale=1},\\&DEmodule\left(F_{in}\right)_{scale=3},\\&DEmodule\left(F_{in}\right)_{scale=5},\\&DEmodule\left(F_{in}\right)_{scale=7})\}\end{aligned} \tag{7}$$

On the other hand, as the scale increases, we can find that the $CP$ of each scale contains all the receptive fields of the previous scale. It can be seen as taking the convolution kernel of the previous scale as a whole as $CP$, and then “Patch” $PSP$ and $EP$ around it.

$$DE_i = Patch\left(DE_{i-1}\right) \quad i > 1 \tag{8}$$

In the process of scale expansion, CP, PSP and EP obtain different degrees of information enhancement, which enhances the contrast of information at the scale level and is more conducive to the network to capture differences in detailed information and achieve higher accurate segmentation.

## 4. Experiments

### 4.1. Datasets

**ISIC (2018)** [44] This dataset, published by the International Skin Imaging Collaborative (ISIC) in conjunction with leading computer vision conferences, is the largest dermoscopic image dataset of its size in the world. The dataset is acquired using a dermatoscope from different sites on patients who underwent skin cancer screening at different institutions. There is a great diversity in the data in terms of color and size. Some of the samples also have the effect of hair masking and boundary blurring, which adds difficulty to the training. For the segmentation task, the dataset contains three parts: (i) a training set comprising 2594 sample images; (ii) a test set with 1000 sample images; and (iii) a validation set consisting of 100 sample images.

**CHAOS-T1** [45] This dataset is derived from CHAOS challenge, the CT images are acquired from the upper abdominal region of the patient after contrast injection. The main challenges contained in this dataset are: adjacent organs having similar Hounsfield value ranges, differences in Hounsfield value ranges for the same tissue, significant shape differences in samples obtained from different patients, and the presence of some atypical liver samples. The dataset contains 647 training samples with a native resolution of 256 × 256 pixels. Since the test set labels in the dataset are not publicly available, in this experiment we use 20% of the original training set for testing and validation, and the remaining part is used for training.

**Clinical Face** The dataset is collected by Shanghai University of Traditional Chinese Medicine, Shanghai, China. The dataset contains a total of 180 human facial images collected in an open environment using the same equipment. The face image samples contain multiple complex backgrounds, such as multiple people in the same frame. Moreover, the lighting, shooting angle and shooting distance vary, which poses a considerable challenge for accurate segmentation. The original data resolution is 1728 × 2592, which is uniformly adjusted to 256 × 256 due to computational limitations. We select 80% of them as the training set, and the remaining 20% is used for validation and testing. Informed consent has been obtained for the publication of identifying images, ethics approval granted by Shanghai Uni TCM Human Research Ethics Committee (approval number:2021-1039-114-01), with data sharing approval.

**FaceImage** This data is collected by the Smart Medicine research team at Wuyi University, Guangdong, China. The dataset is collected using a quadruple diagnostic instrument and contains a total of 755 standard human face examination image samples. The background environment of this facial diagnosis image sample is simple, focusing on the human face, and the main purpose is to improve the research sample for the image study of TCM facial diagnosis. The resolution of the original dataset is 1728 × 2592, which we uniformly adjusted to 256 × 256 to reduce the computational cost. The original data are divided into 80% and 20%, the first part is used for training and the second part is used for validation and testing. Human Research Ethics Committee Ethics approval granted (approval number: [2019]18), this dataset has been previously studied in different areas and relevant publications listed in references.

### 4.2. Segmented evaluation metrics

To quantitatively demonstrate and evaluate the performance of the network model, five segmentation evaluation metrics are used in this paper, including $mIou$ (mean Intersection over Union), $mF1$ (mean F1 score), $PA$ (Pixel Accuracy), $HD$ (Hausdorff Distance) and $mPre$ (mean Precision). $mIou$, $mF1$, $PA$ and $mPre$ all indicate the overall similarity between the predicted result and the ground truth, and larger values indicate higher similarity. $HD$ is more sensitive to the segmented boundary, and smaller values indicate higher similarity of the edge.

$$mIoU = \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k}p_{ij}+\sum_{j=1}^{k}p_{ji}-p_{ii}} \tag{9}$$

$$mF1 = \frac{1}{k}\sum_{i=1}^{k}\frac{2\times p_{ii}}{\sum_{j=1}^{k}p_{ij}+\sum_{j=1}^{k}p_{ji}} \tag{10}$$

$$PA = \frac{\sum_{i=1}^{k}p_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k}p_{ij}} \tag{11}$$

$$mPre = \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k}p_{ij}} \tag{12}$$

$$HD = \max\left(\max_{a\in A}\left\{\min_{b\in B}\|a-b\|\right\}, \max_{b\in B}\left\{\min_{a\in A}\|b-a\|\right\}\right) \tag{13}$$

**Table 1**
Experimental results (mean ± standard deviation) on ISIC (2018).

| Network | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|
| FCN [15] | 72.97 ± 0.61 | 84.37 ± 0.44 | 91.40 ± 0.59 | 85.86 ± .010 | 5.46 ± 0.20 |
| Unet [3] | 74.81 ± 0.45 | 85.59 ± 0.35 | 91.96 ± 0.43 | 85.91 ± 1.37 | 5.43 ± 0.14 |
| Unext [7] | 77.07 ± 0.39 | 87.05 ± 0.53 | 92.77 ± 0.87 | 87.35 ± 0.96 | 5.20 ± 0.15 |
| Att-Unet [8] | 75.42 ± 0.77 | 85.99 ± 0.70 | 92.13 ± 0.91 | 85.70 ± 1.64 | 5.30 ± 0.23 |
| MSNet [46] | 76.76 ± 0.43 | 86.85 ± 0.45 | 92.68 ± 0.63 | 87.34 ± 0.76 | 5.20 ± 0.12 |
| DANet [47] | 76.84 ± 0.57 | 87.02 ± 0.51 | 92.60 ± 0.94 | 85.51 ± 1.23 | 5.38 ± 0.11 |
| EGE-UNet [48] | **78.57 ± 0.38** | 87.25 ± 0.61 | 92.12 ± 0.63 | **87.45 ± 0.85** | 5.21 ± 0.14 |
| AMSUnet [49] | 76.96 ± 0.53 | 85.68 ± 0.71 | 91.58 ± 0.65 | 86.85 ± 1.12 | 5.33 ± 0.14 |
| **MSDEnet (ours)** | 78.49 ± 0.52 | **87.95 ± 0.38** | **93.20 ± 0.37** | 87.35 ± 0.87 | **5.08 ± 0.11** |
| Relative gains[a] | −0.08 | 0.70 | 0.43 | −0.10 | 0.12 |

[a] The smaller the HD value, the better the performance, and the larger the better for all other metrics.

where $p_{ij}$ denotes the number of pixels for which the true category $i$ is predicted to be $j$ and is the number of segmented target categories. $\| \cdot \|$ denotes the distance paradigm between the pixel sets $A$ and $B$. In this paper, the Euclidean distance is used.

### 4.3. Implementation details

The model used in this paper was based on the PyTorch framework and the hardware condition is a single Nvidia RTX A5000. In the training process, the input sample size was adjusted to 256 × 256 and the mini-batch size was 8. For ISIC (2018) and CHAOS-T1, Random horizontal flip, random vertical flip and random 90-degree rotation were used as data enhancement to avoid overfitting. While for Clinical Face and FaceImage, the data enhancement was changed to 15-degree random rotation and random horizontal flip. Adam with decoupled decay (AdamW) was selected as the optimizer. The initial learning rate and weight decay were set as 0.001 and 0.0005, respectively. Warm-up and CosineAnnealingLR was used as the learning strategy and adjust the learning rate according to epoch. Since different models cannot converge at the same rate on the same dataset, the epoch is not set uniformly. Except for epoch, the above parameters are kept consistent during all model training.

### 4.4. Experiments results

To ensure the rigor of the experimental comparison, the training parameters and data sets of all network structures are kept consistent. The experimental models for the comparison are derived from open source code, and all models are retrained.

We experimentally compare the MSDEnet proposed in this paper with currently use medical image segmentation models, which include FCN, Unet, ATT-Unet, Unext, MSNet, DANet, EGE-UNet and AMSUnet. The experimental quantitative results are shown in Tables 1, 3, 5, and 7, and optimal results are indicated by bold. It can be seen that MSDEnet out of a total of 20 metrics on all 4 datasets 15 optimal performances are obtained. The qualitative presentation of the experimental results is shown in Fig. 7.

The data in Tables 1, 3, 5, and 7 represent absolute gains. Relative gains, on the other hand, refer to the relative gap between the performance of MSDEnet and the optimal performance, as well as the relative difference gain to the sub-optimal performance. In Tables 2, 4, 6 and 8, the minimum, median and maximum values were selected from all the experimental results in this dataset after removing outliers. As shown in Fig. 6, we plotted violin boxes for the experimental results on the two publicly available datasets, which allows for a quantitative and intuitive comparison of model performance. It is clear to see that our model performs optimally on several metrics.

Compared with the classical Unet, we have achieved a very excellent performance in all data, which show that the multi-scale capability of CMS's backbone enhancement and MSDE's detail enhancement has significantly improved the segmentation ability of the model. Although Unet++ perceptual field fusion at different scales is performed, the negative effect from excessive feature information fusion makes it not perform well. The attention structure of Att-Unet and DANet can effectively suppress the activation in irrelevant regions, reduce the effect of redundant information, and significantly improve the segmentation results. UNext designs the lightweight and efficient Tok-MLP, which adds local information to the model by window-based attention. However, compared with MSDEnet, the single-scale structure of Att-Unet, DANet and UNxet make them less sensitive to detailed information and less capable of cpturing information of similar colors and fuzzy boundaries to accomplish accurate boundary definition. MSNet proposes a simple and efficient multi-scale subtraction unit (MSU) and uses pyramidal stacking to use features at different levels, and finally obtains rich multi-scale information. Whereas, MSDEnet reduces the generation of redundant feature information and increases the ERF at the source of redundant information generation, i.e., decoding process, which makes MSDEnet obtain better performance compared to MSNet.

**Table 2**
Maximum, median and minimum value on ISIC (2018).

| Network | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|
| Max | 80.11 | 88.54 | 95.30 | 89.24 | 5.83 |
| Med | 76.74 | 86.55 | 92.15 | 86.73 | 5.28 |
| Min | 71.79 | 83.47 | 89.98 | 81.70 | 4.79 |

The performance comparison between MSDEnet and EGE-UNet on ISIC (2018) and CHAOS-T1 indicates minimal differences. This similarity is attributed to both networks incorporating an optimized multi-scale structure, proving effective in boundary resolution. However, on the Clinical Face and FaceImage datasets, MSDEnet has better performance. EGE-UNet is optimized by Group multi-axis Hadamard Product Attention module and Group Aggregation Bridge module, which utilizes lesion information from multiple views and various types of feature information fusion masks to obtain more complex and effective feature information. MSDEnet, on the other hand, utilizes CMS to reduce the dilution of boundary detail information by feature redundancy information in the channel. On this basis, MSDE is utilized for dimensionally extended peripheral information to compare with central information, which makes MSDEnet more sensitive to changes in texture, color, and location information.

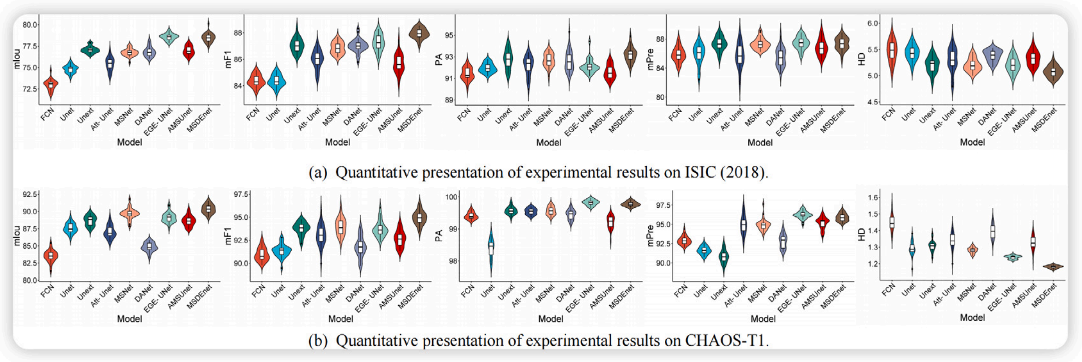
### 4.5. Contrast and ablation study

The results of the various comparative ablation experiments are shown in Tables 9 and 10. We test each structure in the MSDEnet to understand the contribution of each structure to the model. We also include the Res2net module and three common multi-scale structures (Inception, ASPP and RFB) for comparison experiments. The advantages of CMS and MSDE are verified by comparing different combinations of the modules.

Firstly, starting from the Original Unet structure, we replace the encoder with Res2net module and CMS to test them separately. The replaced structure improve in all metrics compared with Original Unet,

**Table 3**
Experimental results (mean ± standard deviation) on CHAOS-T1.

| Network | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|
| FCN [15] | 83.51 ± 0.85 | 90.97 ± 0.71 | 99.41 ± 0.11 | 92.92 ± 0.55 | 1.45 ± 0.20 |
| Unet [3] | 87.58 ± 0.71 | 91.35 ± 0.69 | 98.40 ± 0.28 | 91.62 ± 0.50 | 1.29 ± 0.15 |
| Unext [7] | 88.65 ± 0.73 | 93.75 ± 0.70 | 99.57 ± 0.12 | 90.82 ± 0.80 | 1.31 ± 0.14 |
| Att-Unet [8] | 87.03 ± 0.91 | 93.04 ± 1.17 | 99.54 ± 0.10 | 94.98 ± 1.20 | 1.34 ± 0.16 |
| MSNet [46] | 89.67 ± 0.79 | 93.99 ± 0.92 | 99.56 ± 0.13 | 94.94 ± 0.75 | 1.28 ± 0.13 |
| DANet [47] | 84.83 ± 0.64 | 91.77 ± 1.11 | 99.45 ± 0.15 | 92.81 ± 0.87 | 1.39 ± 0.18 |
| EGE-UNet [48] | 89.11 ± 0.72 | 93.74 ± 0.86 | **99.82 ± 0.08** | **96.22 ± 0.49** | 1.24 ± 0.11 |
| AMSUnet [49] | 88.62 ± 0.75 | 92.68 ± 0.87 | 99.21 ± 0.19 | 95.18 ± 0.75 | 1.33 ± 0.19 |
| **MSDEnet (ours)** | **90.35 ± 0.65** | **94.94 ± 0.74** | 99.78 ± 0.07 | 95.90 ± 0.49 | **1.18 ± 0.11** |
| Relative gains | 0.68 | 1.20 | −0.04 | −0.32 | 0.06 |



(a) Quantitative presentation of experimental results on ISIC (2018).

(b) Quantitative presentation of experimental results on CHAOS-T1.

**Fig. 6.** Quantitative presentation of experimental results.

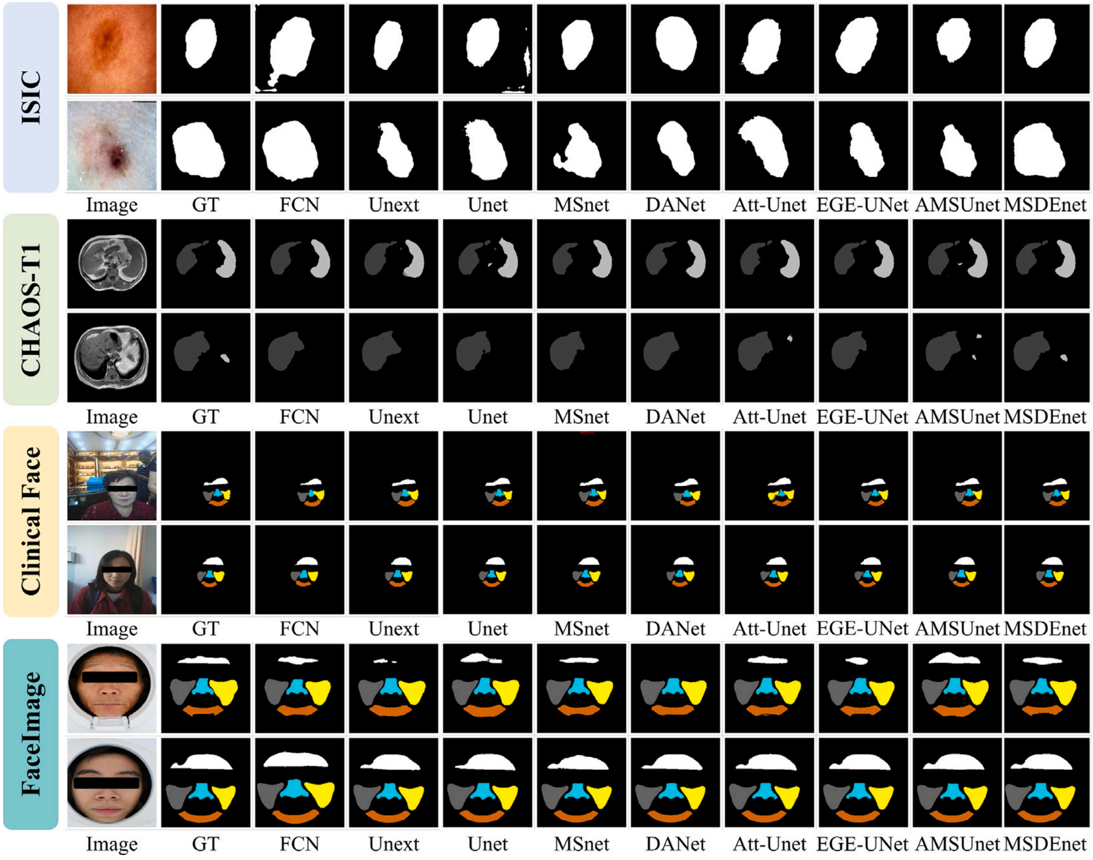


**Fig. 7.** Qualitative comparison of segmentation results.

and the gap between them is not large. Next, we compare the effect of the combination of the Res2net module and the four multi-scale modules. The model performance is improved after the introduction of the multi-scale structure, but the improvement is not obvious, and the performance of the four combinations is basically not much different. Finally, we compare the effect of the combination of CMS and the

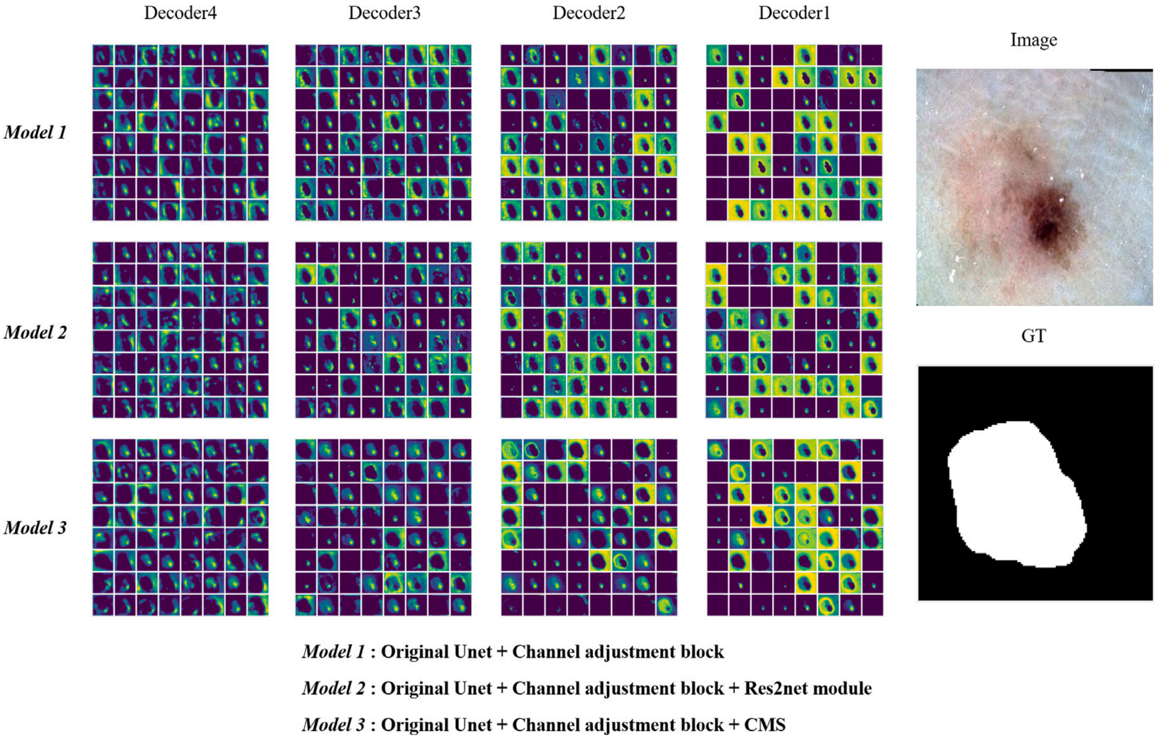

Decoder4
Decoder3
Decoder2
Decoder1
Image
GT
Model 1
Model 2
Model 3
Model 1 : Original Unet + Channel adjustment block
Model 2 : Original Unet + Channel adjustment block + Res2net module
Model 3 : Original Unet + Channel adjustment block + CMS

**Fig. 8.** Visualization of decoder output feature maps.

**Table 4**
Maximum, median and minimum value on CHAOS-T1.

| Network | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|
| Max | 91.76 | 96.52 | 99.93 | 97.64 | 1.56 |
| Med | 88.25 | 93.03 | 99.01 | 94.20 | 1.30 |
| Min | 81.33 | 89.48 | 97.87 | 89.16 | 1.16 |

four multi-scale modules. It is obvious that the CMS brings more improvement compared to the combination of Res2net module and the four multi-scale structures. We believe that the combination of the Res2net module and the four multi-scale structures does not bring significant improvement because the redundant information generated by the Res2net module is more to the CMS, and the frequent addition or concatenation in the multi-scale structure amplifies this factor, and the detailed boundary information is diluted. While independent multi-scale branching of the CMS reduces the generation of redundant information and the introduction of large convolution kernels brings a larger ERF, which is beneficial for segmentation tasks.

## 5. Discussion

### 5.1. Encoder

In order to reduce the generation of redundant information during feature extraction and to increase the ERF of the encoder backbone, the CMS is designed as encoder. Fig. 8 shows the output feature maps from decoder of different model, each model contains different encoder backbone. Since the encoder produces relatively large number of feature map channels, usually 256, 512, 1024 or even 2048, it is difficult to show all comparisons. The difference of the encoder backbone is presented, and the comparison of the output feature maps at the decoder after changing the encoder are shown. To facilitate the presentation, we insert a Channel adjustment block on the jump connection of Original Unet, through which the number of feature map channels passed into the decoder by the three models is guaranteed to be 64, and the output channel of each decoder are equal to the input channels.

In order to visualize the effect of redundant information and multi-scale structure on detail segmentation, we visualized the channel feature maps for different decoders. The closer the black part is to the ground truth, the more accurate the segmentation effect is. From Fig. 8, it can be seen that the segmentation effect of different channel feature maps is better with the increase of the number of decoder layers. At the same time, we can clearly see that Model 1 almost only segments the darker regions, and cannot segment the less obvious regions accurately; Model 2 is able to segment the lighter regions compared to Model 1, but the segmentation accuracy is not high; Model 3 compares with the other two models, the boundary recognition of the lighter regions is more accurate, and the segmentation results are closer to the ground truth. Model 3 is more accurate in identifying the boundary of the light-colored region and the segmentation result is closer to the ground truth. While Model 1 can accurately segment regions with distinct color differences and clear boundaries, it struggles with accurately segmenting regions with colors similar to skin and blurred boundaries. Model 2 benefits from the incorporation of multi-scale structure, as the model becomes more sensitive to detailed information and can capture more subtle changes, leading to better segmentation results than Model 1. Compared to Model 2, Model 3 reduces the influence of redundant information and enables more accurate boundary segmentation. Additionally, the use of large scale convolution kernel expands the ERF for the model, further enhancing its segmentation performance.

### 5.2. Multi-scale structure

Unet-shape networks utilize jump connections to reduce the semantic gap between different depth feature maps, thereby achieving higher quality prediction results. Typically, shallow feature maps contain more detailed information. By incorporating multi-scale structures on the jump connections, it is possible to extract even more detailed feature information. In Fig. 9, three commonly used multi-scale structures are compared with MSDE and present the output results visually. It is

**Table 5**
Experimental results (mean ± standard deviation) on Clinical Face.

| Network | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|
| FCN [15] | 77.69 ± 1.76 | 87.40 ± 1.00 | 98.60 ± 0.08 | 88.96 ± 0.68 | 3.18 ± 0.27 |
| Unet [3] | 82.74 ± 0.77 | 90.48 ± 0.47 | 98.98 ± 0.08 | 91.01 ± 0.89 | 2.85 ± 0.13 |
| Unext [7] | 83.01 ± 0.85 | 90.70 ± 0.43 | 98.93 ± 0.11 | 91.69 ± 1.08 | 2.88 ± 0.16 |
| Att-Unet [8] | 84.57 ± 0.82 | 91.61 ± 0.94 | 99.07 ± 0.06 | 91.72 ± 0.75 | 2.80 ± 0.19 |
| MSNet [46] | 82.91 ± 0.89 | 90.63 ± 0.56 | 98.95 ± 0.10 | 91.09 ± 0.92 | 2.96 ± 0.14 |
| DANet [47] | 79.93 ± 1.08 | 88.83 ± 0.41 | 98.75 ± 0.11 | 89.78 ± 0.60 | 3.07 ± 0.20 |
| EGE-UNet [48] | 83.98 ± 0.79 | 90.52 ± 0.48 | 97.18 ± 0.08 | 90.42 ± 0.79 | **2.79 ± 0.18** |
| AMSUnet [49] | 83.31 ± 0.83 | 89.25 ± 0.61 | 97.92 ± 0.11 | 89.45 ± 1.16 | 3.09 ± 0.26 |
| **MSDEnet (ours)** | **85.79 ± 0.85** | **92.33 ± 0.59** | **99.10 ± 0.07** | **92.74 ± 0.66** | 2.82 ± 0.18 |
| Relative gains | 1.22 | 0.72 | 0.03 | 1.02 | −0.03 |

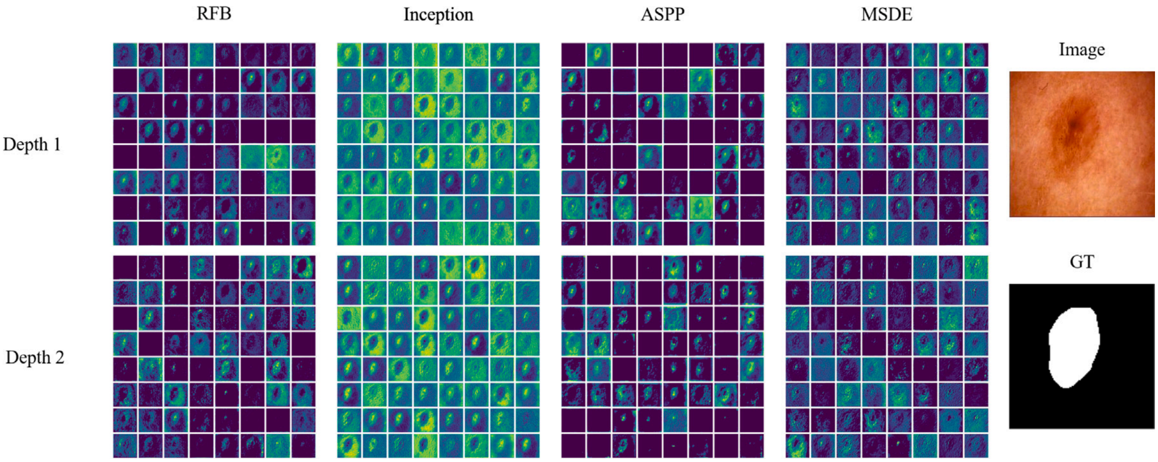


**Fig. 9.** Comparison of multi-scale structure detail extraction capability.

**Table 6**
Maximum, median and minimum value on ISIC Clinical Face.

| Network | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|
| Max | 87.74 | 93.73 | 99.27 | 94.25 | 3.68 |
| Med | 83.26 | 90.54 | 98.88 | 90.87 | 2.91 |
| Min | 72.59 | 85.53 | 97.76 | 87.36 | 2.42 |

evident that the output results of Inception and MSDE contain more detailed feature information when compared to ASPP and RFB, which have a larger number of zero values (represented by black pixels). Further analysis indicates that the output feature maps of Inception are quite similar and less rich in detailed information. In contrast, the output feature map of MSDE contains a diverse range of feature information and exhibits better ability to extract boundary and texture information.

## 6. Conclusions

A simple and effective multi-scale detail enhanced network (MSDEnet) was constructed in order to achieve more accurate medical image segmentation. To address the problem of redundant information being generated in the decoding process, targeted adjustments were made to the Res2net module, allowing the model to perform multi-scale operations independently. The DE inspired by human visual characteristics was then designed, and this enhanced the sensitivity to texture information by superimposing the asymmetric convolution and standard convolution to increase the convolution skeleton and form a distribution with a strong center and weak periphery. We further extended the DE to more scales and designed MSDE. This structure can be viewed as an exploration process of comparing detailed information at different scales, where the central receptive field of each scale contains all the receptive fields of the previous scale. We experimentally compared MSDEnet with commonly used medical image segmentation models on four medical image datasets, and the results show that the model proposed in this paper outperforms other models.

## CRediT authorship contribution statement

**Yuangang Ma:** Writing – original draft, Methodology, Writing – review & editing, Conceptualization, Data curation, Formal analysis. **Hong Xu:** Project administration, Writing – review & editing, Writing – original draft, Funding acquisition, Resources, Supervision. **Yue Feng:** Project administration, Writing – original draft, Writing – review & editing, Funding acquisition, Supervision, Resources. **Zhuosheng Lin:** Project administration, Writing – original draft, Writing – review & editing, Supervision. **Fufeng Li:** Investigation, Data curation. **Xin Wu:** Writing – original draft, Writing – review & editing, Validation. **Qichao Liu:** Writing – original draft, Writing – review & editing, Validation. **Shuangsheng Zhang:** Investigation, Data curation.

## Declaration of competing interest

All authors disclosed no relevant relationships.

## Acknowledgments

This work is supported by the Basic Research and Applied Basic Research Key Project in General Colleges and Universities of Guangdong Province, China (Grant No. 2021ZDZX1032); the Special Project

**Table 7**
Experimental results (mean ± standard deviation) on FaceImage.

| Network | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|
| FCN [15] | 87.60 ± 0.47 | 92.81 ± 0.34 | 95.72 ± 0.47 | 92.72 ± 0.49 | 3.93 ± 0.47 |
| Unet [3] | 88.53 ± 0.36 | 93.90 ± 0.34 | 96.72 ± 0.35 | 91.65 ± 0.82 | 3.90 ± 0.36 |
| Unext [7] | 88.87 ± 0.47 | 94.10 ± 0.29 | 96.91 ± 0.40 | 92.60 ± 0.52 | 3.82 ± 0.29 |
| Att-Unet [8] | 89.42 ± 0.34 | 94.41 ± 0.26 | 97.16 ± 0.45 | 94.23 ± 0.60 | 3.82 ± 0.30 |
| MSNet [46] | 89.60 ± 0.28 | 94.51 ± 0.23 | 97.15 ± 0.55 | 94.47 ± 0.57 | 3.78 ± 0.22 |
| DANet [47] | 88.88 ± 0.40 | 94.11 ± 0.26 | 96.96 ± 0.39 | 93.95 ± 0.76 | 3.87 ± 0.25 |
| EGE-UNet [48] | 88.76 ± 0.36 | 93.72 ± 0.24 | 96.38 ± 0.32 | 93.85 ± 0.46 | 3.88 ± 0.31 |
| AMSUnet [49] | 87.16 ± 0.64 | 93.75 ± 0.30 | 95.87 ± 0.42 | 94.32 ± 0.56 | 3.89 ± 0.43 |
| **MSDEnet (ours)** | **90.82 ± 0.31** | **94.69 ± 0.18** | **97.21 ± 0.28** | **95.57 ± 0.41** | **3.74 ± 0.31** |
| Relative gains | 1.22 | 0.18 | 0.05 | 1.10 | 0.04 |

**Table 8**
Maximum, median and minimum value on FaceImage.

| Network | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|
| Max | 91.26 | 95.18 | 98.51 | 99.06 | 4.90 |
| Med | 88.84 | 94.22 | 96.72 | 94.55 | 3.84 |
| Min | 85.89 | 92.15 | 94.35 | 90.23 | 3.00 |

**Table 9**
Contrast and ablation experiments on ISIC (2018).

| Network | Encoder | Skip connection | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|---|---|
| Original Une | 3 × 3conv | / | 74.81 ± 0.07 | 85.59 ± 0.05 | 91.96 ± 0.03 | 85.91 ± 0.13 | 5.43 ± 0.02 |
| | Res2net | / | 76.34 ± 0.05 | 86.58 ± 0.03 | 92.44 ± 0.12 | 86.07 ± 1.24 | 5.31 ± 0.03 |
| | CMS | / | 76.24 ± 0.09 | 86.80 ± 0.44 | 92.77 ± 0.01 | 85.72 ± 0.88 | 5.29 ± 0.03 |
| MSDEne | Res2net | ASPP | 76.62 ± 0.66 | 86.76 ± 0.42 | 92.68 ± 0.33 | 87.99 ± 1.93 | 5.19 ± 0.09 |
| | Res2net | Inception | 76.37 ± 0.18 | 86.75 ± 0.29 | 92.58 ± 0.07 | 86.67 ± 1.90 | 5.24 ± 0.04 |
| | Res2net | RFB | 76.68 ± 0.49 | 86.80 ± 0.31 | 92.68 ± 0.25 | 87.70 ± 1.56 | 5.18 ± 0.03 |
| | Res2net | MSDE | 77.18 ± 0.18 | 86.90 ± 0.12 | 92.68 ± 0.15 | 88.55 ± 1.78 | 5.16 ± 0.03 |
| | CMS | ASPP | 77.07 ± 0.10 | 86.93 ± 0.07 | **93.26 ± 0.17** | 87.63 ± 1.73 | 5.12 ± 0.08 |
| | CMS | Inception | 76.96 ± 0.38 | 87.28 ± 0.24 | 92.76 ± 0.20 | 87.61 ± 1.65 | 5.17 ± 0.07 |
| | CMS | RFB | 76.98 ± 0.19 | 87.20 ± 0.15 | 92.83 ± 0.16 | **88.76 ± 1.20** | 5.16 ± 0.05 |
| | CMS | MSDE | **78.49 ± 0.19** | **87.95 ± 0.12** | 93.20 ± 0.12 | 87.35 ± 1.01 | **5.08 ± 0.02** |

**Table 10**
Contrast and ablation experiments on CHAOS-T1.

| Network | Encoder | Skip connection | mIou | mF1 | PA | mPre | HD |
|---|---|---|---|---|---|---|---|
| Original Une | 3 × 3conv | / | 87.58 ± 0.26 | 91.35 ± 0.25 | 98.40 ± 0.13 | 91.62 ± 0.31 | 1.29 ± 0.02 |
| | Res2net | / | 88.17 ± 0.02 | 93.70 ± 0.01 | 99.56 ± 0.01 | 94.55 ± 0.23 | 1.30 ± 0.01 |
| | CMS | / | 88.18 ± 0.03 | 93.72 ± 0.09 | 99.55 ± 0.02 | 93.81 ± 0.33 | 1.29 ± 0.02 |
| MSDEne | Res2net | ASPP | 88.54 ± 0.04 | 93.91 ± 0.02 | 99.57 ± 0.08 | 93.58 ± 0.85 | 1.31 ± 0.03 |
| | Res2net | Inception | 88.59 ± 0.06 | 93.94 ± 0.03 | 99.56 ± 0.01 | 94.80 ± 0.17 | 1.31 ± 0.03 |
| | Res2net | RFB | 88.64 ± 0.15 | 93.97 ± 0.08 | 99.58 ± 0.01 | 94.49 ± 0.37 | 1.31 ± 0.01 |
| | Res2net | MSDE | 88.77 ± 0.03 | 94.04 ± 0.02 | 99.58 ± 0.01 | 94.63 ± 0.29 | 1.31 ± 0.02 |
| | CMS | ASPP | 89.37 ± 0.12 | 94.38 ± 0.07 | 99.60 ± 0.01 | 94.98 ± 0.12 | 1.27 ± 0.01 |
| | CMS | Inception | 89.62 ± 0.34 | 94.52 ± 0.19 | 99.61 ± 0.01 | 94.99 ± 0.33 | 1.26 ± 0.01 |
| | CMS | RFB | 89.47 ± 0.39 | 94.48 ± 0.29 | 99.67 ± 0.09 | 95.21 ± 0.18 | 1.25 ± 0.03 |
| | CMS | MSDE | **90.35 ± 0.23** | **94.94 ± 0.12** | **99.78 ± 0.03** | **95.90 ± 0.40** | **1.18 ± 0.02** |

of Guangdong Province, China (Grant No. 2020A1313030021); and the Scientific Research Project of Wuyi University (Grants No. 2018TP023 and 2018GR003).

## References

[1] Hassan Homayoun, Hossein Ebrahimpour Komleh, Automated segmentation of abnormal tissues in medical images, J. Biomed. Phys. Eng. 11 (2019) 415–424.

[2] Gobert Lee, Hiroshi Fujita, Deep Learning in Medical Image Analysis: Challenges and Applications, Vol. 1213, 2020.

[3] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, 2015, pp. 234–241.

[4] Xiaoqi Zhao, Hongpeng Jia, Youwei Pang, et al., M$^2$SNet: Multi-scale in multiscale subtraction network for medical image segmentation, 2023, arXiv preprint arXiv:2303.10894.

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, et al., TransUNet: Transformers make strong encoders for medical image segmentation, 2021, arXiv abs/2102.04306.

[6] Yunhe Gao, Mu Zhou, Dimitris N. Metaxas, Utnet: a hybrid transformer architecture for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, 2021, pp. 61–71.

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

[8] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, M.J. Lee, et al., Attention U-net: Learning where to look for the pancreas, 2018, arXiv abs/1804.03999.

[9] Debesh Jha, Pia Helen Smedsrud, M. Riegler, et al., ResUNet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia, ISM, 2019, pp. 225–2255.

[10] Run Su, Deyun Zhang, Jinhuai Liu, Chuandong Cheng, MSU-net: Multi-scale U-net for 2D medical image segmentation, Front. Genet. 12 (2021) 639930.

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, et al., Pyramid scene parsing network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 6230–6239.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, et al., Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2014, pp. 1–9.

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, et al., Rethinking atrous convolution for semantic image segmentation, 2017, arXiv abs/1706.05587.

[14] Yanghao Li, Yuntao Chen, Naiyan Wang, et al., Scale-aware trident networks for object detection, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 6053–6062.

[15] Evan Shelhamer, Jonathan Long, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2014, pp. 3431–3440.

[16] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, et al., Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision–MICCAI 2018, 2018, pp. 3–11.

[17] Sahar Yousefi, Hessam Sokooti, Mohamed S. Elmahdy, et al., Esophageal tumor segmentation in CT images using a dilated dense attention unet (ddaunet), IEEE Access 9 (2020) 99235–99248.

[18] Yutong Cai, Yong Wang, Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation, in: Third International Conference on Electronics and Communication; Network and Computer Technology, ECNCT 2021, Vol. 12167, 2022, pp. 205–211.

[19] Edwin Thomas, S.J. Pawan, Shushant Kumar, et al., Multi-res-attention UNet: A CNN model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images, IEEE J. Biomed. Health Inf. 25 (2020) 1724–1734.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, 2020, arXiv abs/2010.11929.

[21] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, et al., Mlp-mixer: An all-mlp architecture for vision, Adv. Neural Inf. Process. Syst. 34 (2021) 24261–24272.

[22] Zhiqin Zhu, Xianyu He, Guanqiu Qi, et al., Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, Inf. Fusion 91 (2023) 376–387.

[23] Xianyu He, Guanqiu Qi, Zhiqin Zhu, et al., Medical image segmentation method based on multi-feature interaction and fusion over cloud computing, Simul. Model. Pract. Theory 126 (2023) 102769.

[24] Yang Xu, Xianyu He, Guofeng Xu, et al., A medical image segmentation method based on multi-dimensional statistical features, Front. Neurosci. 16 (2022).

[25] Yuanyuan Li, Ziyu Wang, Li Yin, et al., X-Net: a dual encoding–decoding method in medical image segmentation, Vis. Comput. (2023).

[26] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, et al., Res2net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2019) 652–662.

[27] Thomas Serre, Lior Wolf, Stanley Bileschi, et al., Robust object recognition with cortex-like mechanisms, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 411–426.

[28] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[29] Maoke Yang, Kun Yu, Chi Zhang, et al., Denseaspp for semantic segmentation in street scenes, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3684–3692.

[30] Songtao Liu, Di Huang, et al., Receptive field block net for accurate and fast object detection, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 385–400.

[31] Xuebin Qin, Zichen Zhang, Chenyang Huang, et al., U2-net: Going deeper with nested U-structure for salient object detection, Pattern Recognit. 106 (2020) 107404.

[32] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, et al., Deeply supervised salient object detection with short connections, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 5300–5309.

[33] Chao Peng, Xiangyu Zhang, Gang Yu, et al., Large kernel matters — Improve semantic segmentation by global convolutional network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 1743–1751.

[34] Jingdong Wang, Ke Sun, Tianheng Cheng, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3349–3364.

[35] Fisher Yu, Vladlen Koltun, Multi-scale context aggregation by dilated convolutions, 2015, CoRR abs/1511.07122.

[36] Fisher Yu, Vladlen Koltun, Thomas A. Funkhouser, Dilated residual networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 636–644.

[37] Nikhil Kumar Tomar, Abhishek Srivastava, Ulas Bagci, et al., Automatic polyp segmentation with multiple kernel dilated convolution network, in: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems, CBMS, 2022, pp. 317–322.

[38] Wenjie Luo, Yujia Li, Raquel Urtasun, et al., Understanding the effective receptive field in deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 29 (2016).

[39] Xiaohan Ding, X. Zhang, Yi Zhou, et al., Scaling up your kernels to 31 × 31: Revisiting large kernel design in CNNs, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 11953–11965.

[40] Jerome Y. Lettvin, On seeing sidelong, Sciences (New York) 16 (1976) 10–20.

[41] Barbara Sivak, Christine L. MacKenzie, Chapter 10 the contributions of peripheral vision and central vision to prehension, Adv. Psychol. 85 (1992) 233–259.

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, et al., Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[43] Xiaohan Ding, Yuchen Guo, Guiguang Ding, et al., ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 1911–1920.

[44] Noel C.F. Codella, David Gutman, M. Emre Celebi, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018, 2018, pp. 168–172.

[45] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, et al., CHAOS challenge-combined (CT-mr) healthy abdominal organ segmentation, Med. Image Anal. 69 (2021) 101950.

[46] Xiaoqi Zhao, Lihe Zhang, Huchuan Lu, Automatic polyp segmentation via multi-scale subtraction network, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, 2021, pp. 120–130.

[47] J. Fu, J. Liu, Haijie Tian, et al., Dual attention network for scene segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 3141–3149.

[48] Jingsheng Gao Jiacheng Ruan, et al., EGE-UNet: an efficient group enhanced unet for skin lesion segmentation, 2023, arXiv:2307.08473.

[49] Yunchou Yin, Zhimeng Han, Muwei Jian, et al., AMSUnet: A neural network using atrous multi-scale convolution for medical image segmentation, Comput. Biol. Med. 162 (2023) 107120.