Mathematical Biosciences and Engineering

MBE, 21(2): 2753–2767.
DOI: 10.3934/mbe.2024122
Received: 13 November 2023
Revised: 01 January 2024
Accepted: 15 January 2024
Published: 24 January 2024

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

Research article

DAU-Net: A medical image segmentation network combining the Hadamard product and dual scale attention gate

Xiaoyan Zhang[1], Mengmeng He[1,*] and Hongan Li[1,2]

[1] School of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China
[2] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

*** Correspondence:** Email: 22208223093@stu.xust.edu.cn; Tel: +8618161707216.

Abstract: Medical image segmentation has an important application value in the modern medical field, it can help doctors accurately locate and analyze the tissue structure, lesion areas, and organ boundaries in the image, which provides key information support for clinical diagnosis and treatment, but there are still a large number of problems in the accuracy of the segmentation, so in this paper, we propose a medical image segmentation network combining the Hadamard product and dual-scale attention gate (DAU-Net). First, the Hadamard product is introduced in the structure of the fifth layer of the codec for element-by-element multiplication, which can generate feature representations with more representational capabilities. Second, in the jump connection module, we propose a dual scale attention gating (DSAG), which can highlight more valuable features and achieve more efficient jump connections. Finally, in the decoder feature structure, the final segmentation result is obtained by aggregating the feature information provided by each part, and decoding is achieved by up-sampling operation. Through experiments on two public datasets, Luna and Isic2017, DAU-Net is able to extract feature information more efficiently using different modules and has better segmentation results compared to classical segmentation models such as U-Net and U-Net++, and also verifies the effectiveness of the model.

Keywords: medical image segmentation; two-scale attention gates; Hadamard product; fully convolutional networks; feature fusion

1. Introduction

Medical image segmentation, as a key task in the field of medical image processing, aims at accurately identifying and separating different tissue structures, organs, and lesion regions, etc. from medical images. With the continuous development of medical imaging technology, the wide application of high-resolution medical images, such as computed tomography (CT) [1,2] and ultrasonography [3], provides physicians with certain diagnostic and therapeutic tools. However, medical images often have highly complex structures, including organs, blood vessels, tumors, various tissue structures, and lesion areas [4–6], and it is a complex and time-consuming task to extract useful information from them and perform accurate analysis. Typically, quality and noise issues in medical images may also affect the accuracy of the segmentation results, thus requiring higher-level algorithms and networks to capture the subtle features of these structures, which makes medical image segmentation even more challenging. To cope with these challenges, researchers continue to propose new methods and algorithms. From traditional image segmentation algorithms [7–11] to image segmentation models based on convolutional neural networks [12–14], many innovative solutions have emerged in the field of medical image segmentation, and in particular, the rapid development of deep learning technology has revolutionized medical image segmentation.

The emergence of convolutional neural networks (CNNs) and the introduction of fully convolutional networks (FCNs) have enabled image segmentation to no longer rely on traditional handcrafted feature engineering [15], but rather to learn and predict from the pixel level end-to-end, which greatly improves segmentation accuracy. However, these architectures face an important challenge in that critical detail information is often lost in the deep structure of the network. To cope with this problem, Ronneberger et al. first proposed the U-Net architecture [16], which employs an encoder and decoder design that allows the model to efficiently capture feature information at different scales, and introduces intermediate signal hopping connections in the symmetric structure to achieve this, which significantly improves performance in medical image segmentation. Inspired by the U-Net architecture in subsequent studies, Oktay et al. [17] proposed Attention U-Net, which aims to introduce the self-attention mechanism into the U-Net architecture so that the model can focus more on the important regions in the image, which helps to better capture the subtle image features and thus improves the accuracy and robustness of the segmentation. Zhou et al. [18] proposed the U-Net++ structure, the core idea of this method is to construct a pyramidal feature extraction network, which significantly improves the performance of the model by introducing a multi-scale feature fusion mechanism to reduce the lost semantic information, enabling the model to better understand the information in the image. Huang et al. [19] proposed the U-Net3+, which is an extension and improvement of the classical U-Net, and designed a full-scale jump-joining method to combine low- and high-resolution information at different scales to further improve segmentation performance. He et al. [20] proposed ResNet by introducing the residual module, which uses jump connections to simplify the training of the deep network and solves the problem of the increase in the number of layers of the neural network that leads to difficulty in training. Alom et al. [21] proposed the R2U-Net model by combining residual networks, U-Net, and recurrent residual neural networks (RCNN), which overcomes the problem of gradient vanishing in deep neural networks and realizes the iterative transfer of information while preserving the contextual information and local details, enabling R2U-Net to perform multiple loop operations on different scales, which helps to better understand the global context of an image. Chen et al. [22] proposed TransUNet, which for the first time is based on the Transformer module, allowing the model to automatically capture global information and long-range dependencies in an image. In addition, TransUNet introduces jump connectivity and deep feature fusion to preserve

local details and contextual information in images. This comprehensive architecture allows TransUNet to excel in medical image segmentation, especially for images with complex structures and rich textures. Ruan et al. [23] proposed EGE-UNet, which combines a Hadamard Product Attention Module (GHPA) and a Group Aggregation Bridge Module (GAB) in a lightweight manner. GHPA groups input features and performs Hadamard Product Attention Mechanism (HPA) on different axes to extract pathology information from different perspectives. GAB effectively fuses multi-scale information by grouping low-level features, high-level features, and masks generated by the decoder at each stage, while still providing high segmentation performance with only 50KB of parameters. Valanarasu et al. [24] proposed UNeXt, which inherits the encoder-decoder structure of U-Net and the design of the hopping connection, retaining the model advantage of processing local features and global contextual information, while one of the key innovations is the introduction of the ability to process multimodal image data. UNeXt is able to process information from multiple modalities simultaneously, fusing them together so as to fully utilize the complementary information of different modalities and to improve the accuracy of segmentation. This cross-modal fusion makes UNeXt suitable for a wider range of medical image segmentation tasks, capturing fine structures and textures in medical images, while possessing strong generalization ability. Tang et al. [25] proposed CMUNeXt, which is a network that improves on the UNeXt model to enable fast and accurate assisted diagnosis in real scene scenarios. CMUNeXt utilizes its large kernel and inverted bottleneck design to thoroughly mix long-range spatial and positional information to efficiently extract global contextual information, in addition to introducing the Skip-Fusion block, which aims to realize smooth jump connections and ensure sufficient feature fusion. In addition to this, we found an attention module based dual encoder decoder for colonoscopic polyp segmentation network PSNet proposed by Lewis et al. [26]. Specifically, this network consists of a dual encoder and decoder, which is composed of a (polyp segmentation)PS encoder, Transformer encoder, PS decoder, Enhanced Expansion Transformer decoder, partial decoder, and merge module are synthesized. This dual codec structure enables efficient feature extraction and exploitation, plus synchronized codec operations help to better capture key features in the image. In addition, PSNet utilizes skip connections to preserve feature information at different levels, and attention mechanisms are incorporated into almost every level and module of the network to further increase the network's focus on important regions. This helps the network to be able to segment polyps in colonoscopy images more accurately.

In this paper, we propose DAU-Net, a medical image segmentation network that combines the Hadamard product and biscale attention gates. At the encoder head, we introduce a stem module to extract raw features from the input image. Extracting enough feature information from the beginning can compensate to some extent for the spatial information lost during subsequent feature extraction operations. In layers 1–4 of the network, we use depth-separable convolution and point-by-point convolution to extract spatial and channel information. In the convolution block of layer 5, we apply the Hadamard product algorithm to fuse the feature maps of different levels or channels to deal with the variation of the image at different scales. In the decoder stage, we propose a combination of Dual Scale Attention Gate (DSAG) and up-sampling operations, which is a module mainly composed of point-by-point convolution and dilation convolution, where point-by-point convolution is mainly used to reduce the dimensionality of the features, and dilation convolution is used to expand the sensory field and capture more contextual information so as to better capture the global and local structures in the image. In summary, the main contributions of this paper can be summarized as follows:

1) The DAU-Net model is proposed, which introduces a stem module in the head of the encoder for extracting the original features of the image to prevent too much feature information from being lost due to the jump connections at the top layer, and applies the Hadamard product operation at the last layer of

the codec, which can multiply the feature information of the feature maps element-by-element, thus generating feature representations with more characterizing ability, which can help to improve the comprehension of the semantic information of the different levels and enhance the accuracy of the segmentation;

2) In the decoder stage, we combine the up-sampling operation with the proposed DSAG, which is a module that mainly employs point-by-point convolution and dilation convolution composition, and then configures the activation function and the batch normalization process to connect these two kinds of convolution for feature-feature fusion and information transfer, and then finally screen out the more valuable features through a convolution with a voting mechanism, which improves the performance and applicability of the model;

3) Experimental analysis using two different datasets of lung CT images and skin lesion segmentation, the experimental results show that the DAU-Net model proposed in this paper is better than the previous SOTA segmentation models in terms of Iou and Dice coefficients, and has better segmentation performance and higher accuracy.

The rest of the paper is organized as follows: Section 2 describes the overall architecture of the network and the detailed aspects of our proposed method. Section 3 introduces the experimental design section and provides experimental results, analyzes the network proposed in this paper in comparison with other classical networks, and verifies the usefulness of the key modules proposed in DAU-Net. In Section 4, we discuss several state-of-the-art networks that are currently available in the segmentation domain based on the attention mechanism and have a lightweight architecture. Finally, the paper is summarized in Section 5.

2. The proposed methods

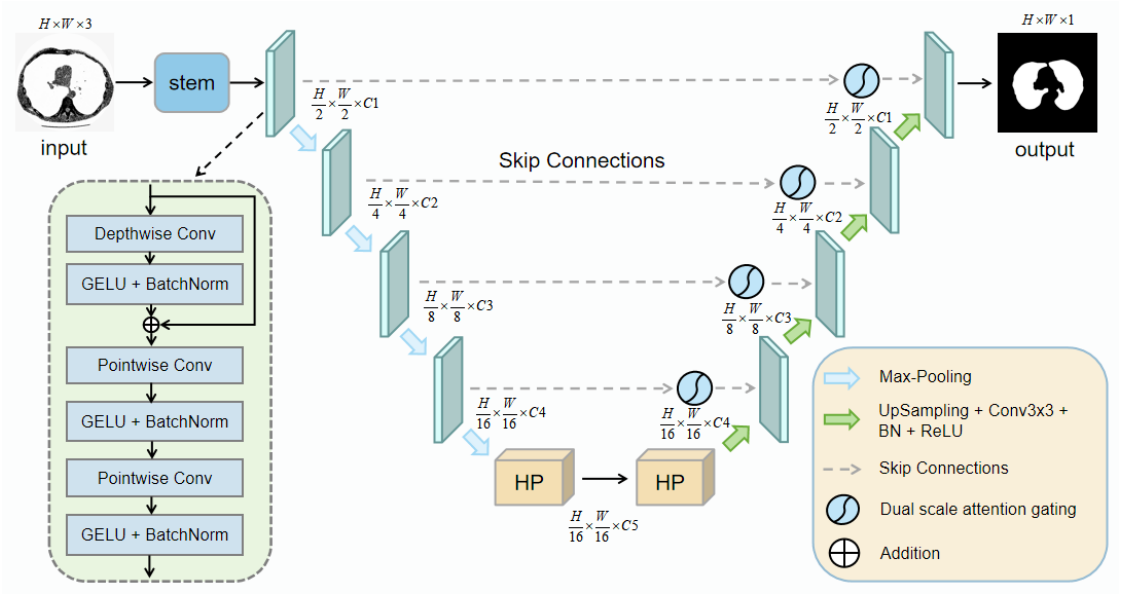

Figure 1. Structure of the DAU-Net network, where the number of channels C1~C5 of the network is set to 32, 64, 128, 256, and 512, respectively.

The network architecture of our proposed DAU-Net is shown in Figure 1. It has a similar codec structure to U-Net, which consists of a total of five layers from top to bottom, and is divided into two stages: encoder and decoder with jump connections. In the encoder stage, semantic feature information

of medical images is extracted by constructing a convolutional neural network, and the number of channels is extended with convolutional blocks. In the decoder stage, the features extracted by dual-scale attention gating are fused with the up-sampled features of the decoder to realize the global semantic information of the medical image is captured and accurately segmented.

2.1. Encoder stage

The encoder has five layers from top to bottom. The first four layers include depth separable convolution, point-by-point convolution, GELU activation function, batch normalization, and downsampling operations. At the fifth layer, we introduce the Hadamard product algorithm in the convolution, and in addition to that, we include a stem block after the input for extracting the original features of the input image. In order to avoid the loss of important semantic feature information due to inconsistent jump connections at the top layer, we set up two ordinary convolutional blocks consisting of a convolutional layer with a convolutional kernel size of 3×3, a step size of 1, a padding of 1, a batch normalization layer, and a ReLU activation function, respectively.

In the encoder section, the main components of the convolution block are depth-separable convolution and point-by-point convolution. Compared to normal convolution, depth-separable convolution effectively reduces memory requirements and reduces the computational effort of convolution operations. In the encoder, we use a deep convolution with a larger convolution kernel size to extract the global information of each channel and then perform residual concatenation. In order to fully fuse the spatial and channel information, we apply two point-by-point convolutions after the deep separable convolution. At the same time, we set the hidden dimension between the two point-by-point convolution layers to be four times the width of the input dimension, and by extending the hidden dimension, we are able to more fully and comprehensively blend the global spatial dimension information extracted by the deep convolution. In addition, we process the paradigm layers using the GELU activation function and batch normalization after each convolution. The encoder part of the convolution block can be defined as

$$f_t^{'} = BN(\sigma_1\{DepthwiseConv(f_{t-1})\}) + f_{t-1} \tag{1}$$

$$f_t^{''} = BN(\sigma_1\{PointwiseConv(f_t^{'})\}) \tag{2}$$

$$f_t = BN(\sigma_1\{PointwiseConv(f_t^{''})\}) \tag{3}$$

where f_t denotes the output feature mapping of layer 1 in the stem block, σ_1 denotes the GELU activation function, and BN denotes the batch normalization layer.

2.2. *The Hadamard product*

In deep convolutional neural networks, the Hadamard product can fuse feature maps of different levels or channels. By multiplying the corresponding elements of two feature maps A and B with the same dimensionality, fusing them element by element can combine the relevant information of the two feature maps. We use this algorithm in the fifth layer of the encoder and decoder structure to generate feature representations with more representational power. This helps to improve the understanding of different levels and semantic information and increase the accuracy of segmentation. The mathematical

expression for the Hadamard product operation is

$$C = A \odot B \tag{4}$$

where C is a new tensor with the same shape as A and B. Each element of C is the product of the elements at the corresponding positions in A and B, i.e., the elements in the i-th row and j-th column of C are equal to the product of the elements in the i-th row and j-th column of A and B.

In this convolution block, given an input x and a randomly initialized learnable tensor p, bilinear interpolation is used to adjust the size of p to match the size of x. Then, we use a depth-separable convolution on p, followed by a hadamard product operation between x and p to obtain the output. Subsequently, after our proposed biscale attention gate, the robustness of the model is improved by subjecting images from different scales or resolutions to hadamard product operations, which helps to deal with the variation of objects in the image at different scales.

2.3. Dual scale attention gating

The decoder stage also consists of a total of five layers from top to bottom. Each layer consists of a jump connection module and an upsampling block, connected in the middle by the dual-scale attention gate(DSAG) proposed in this paper, as shown in Figure 2. In the DSAG, in order to adaptively select semantic features with different resolutions, we use a point-by-point convolution and a dilation convolution to extract semantic features with different receptive fields. Among them, the point-by-point convolution is mainly used to reduce the dimensionality of the features, and the dilation convolution is used to expand the receptive field and capture more contextual information. Each convolution is equipped with a GELU function and a batch normalization layer for feature fusion and information transfer, and these two convolutions are connected to generate feature maps of the same size. Subsequently, the output feature maps are then connected and the most valuable features are selected by a convolutional block with a voting mechanism, and finally a Sigmoid activation function is connected to control the output between 0–1. The module can be defined as

$$\begin{aligned} f_{Concat} = Concat(&BN\{\sigma_2\{PointwiseConv(f)\}\}, \\ &BN\{\sigma_2\{DilationConv(f)\}\}) \end{aligned} \tag{5}$$

$$f_m = f \times \sigma_3(VoteConv(f_{Concat})) + f \tag{6}$$

where f_{Concat} denotes connecting the features, f_m comes from the output features of the DSAG, f represents the encoded features, and σ_2 and σ_3 denote the GELU and Sigmoid activations, respectively.

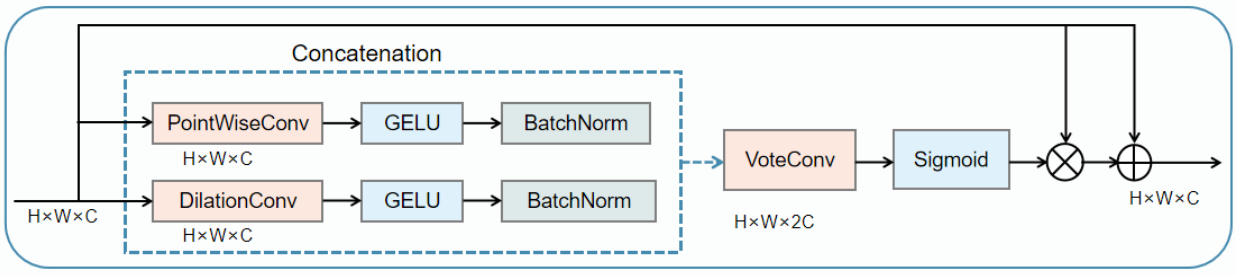

Figure 2. Dual scale attention gating.

2.4. Decoder stages with skip-connections

The decoder part is also composed of five layers from bottom to top, each layer including a skip-connection and an upsampling module. Most of the traditional skip-connections use ordinary convolution operations for feature fusion. In our encoder part, we set up a Group convolution with a Group of 2 to extract the features obtained from the skip-connections and up-sampling operations one by one, respectively, and the convolution kernel size of this convolution is set to be 3×3, with a step size of 1 and a padding of 1. In order to merge the extracted features enough to be fused, we merge two inverse point-by-point convolutions after the Group convolution, which adaptively assigns the features before fusion to the Group convolution and does a large amount of feature fusion. Each convolution is followed by a GELU activation function and a BatchNorm layer as defined below:

$$f_n = Concat(BN\{OridinaryConv(f_\varepsilon)\}, BN\{OridinaryConv(f_\eta)\}) \tag{7}$$

$$f_x = BN(\sigma_1\{PointwiseConv(f_n)\}) \tag{8}$$

$$f_x^{'} = BN(\sigma_1\{PointwiseConv(f_x)\}) \tag{9}$$

Ultimately, $f_x^{'}$ denotes the output fused feature map in the decoder, while f_ε and f_η denote the features obtained by the jump-join and up-sampling operations, respectively. The purpose of this series of operations is to fuse the extracted features in order to produce a feature representation with rich information that facilitates better performance of the model in the task.

3. Experiments and result analysis

3.1. Experiment environment and parameter configuration

The programming language used in this experiment is Python 3.8, and the deep learning framework is Pytorch 1.7, CUDA version 11.1. The processor is the Intel(R) Xeon(R) Platinum 8255C, the graphics card is the NVIDIA RTX 2080Ti discrete graphics card, and 40G of RAM, there is the training and testing are carried out on a Linux operating system. In the training process, we use the binary cross-entropy loss BceLoss, the batch size is set to 4, and the number of epochs is set to 200 times during network training.

3.2. Datasets

In order to verify the effectiveness of the model, this paper uses 2 different types of public datasets in the medical field. The datasets are medical image segmentation tasks in different modalities, which are lung segmentation and skin lesion segmentation in CT images. The lung segmentation dataset Luna dataset in CT images is from the Kaggle Lung Nodule Analysis Competition in 2017. In the Luna dataset for lung data, a total of 267 images were included, of which 213 images were used as the training set, 54 images were used as the test set, and the validation set was the same as the test set. The skin lesion segmentation dataset isic2017 is provided by the International Skin Imaging Collaboration (ISIC). The ISIC 2017 dataset has three sets: the training set (2000 images), the validation set (150

images), and the test set (600 images), with a total of 2750 images. Ground truth values for the mask images in these datasets were generated using various techniques. All data were reviewed and organized by practicing dermatologists with expertise in dermoscopy, which is accurate and reliable. The image data split for the training set, validation set, and test set is shown in Table 1.

Table 1. Split segmentation of image data for training set, validation set, and test set.

Dataset	Train	Valid	Test	Total
Luna	213	-	54	267
Isic2017	2000	150	600	2750

3.3. Evaluation index

The evaluation indexes in this paper mainly consist of the intersection and merger ratio (Iou), Dice similarity coefficient, Hd (Hausdorff distance, Hd) coefficient, and cross entropy loss function (Binary Cross Entropy Loss, BCELoss) to comprehensively evaluate the performances of different models. At the same time, we also include the processing time of each network and the number of learnable parameters to show the performance of each network comprehensively.

The Iou and the Dice coefficient are evaluation metrics commonly used in medical image segmentation, which take into account the degree of spatial overlap between model prediction and actual annotation, and are more suitable for facing pixel-level matching in segmentation tasks. In this paper, Iou denotes the degree of area similarity between the segmented object and the original object, and Dice coefficient measures the two ensemble similarity metrics, both taking values in the range of [0, 1]. In Eqs (10) and (11), A is the prediction result and B is the real labeling value; the larger the value of the indicator, the higher the similarity with the actual results, the better the segmentation results.

Hd is a measure that describes the degree of similarity between the two sets of point sets, which indicates that the segmentation results and the labeling results of the two sets of point sets between the shortest distance of the maximum value measure the maximum degree of mismatch between the two. In Eq (12), h(A,B) and h(B,A) are the one-way Hausdorff distance from set A to set B and from set B to set A, respectively. h(A,B) first ranks the distance between each point a_i in the set of points A to point b_j in set B which is closest to this point, and finally takes the maximum value of this distance as h(A,B); the smaller the value means the closer the segmentation result and the labeling result. BCELoss has good effect on the multiclassification image segmentation problem. In Eq (13), N is the total number of medical image samples, y_i is the category to which the ith sample belongs, and p_i is the predicted value of the ith sample, and the smaller the value of loss means the closer the model segmentation result and the real labeling result.

$$Iou(A,B) = \frac{A \cap B}{A \cup B} \tag{10}$$

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{11}$$

$$H(A,B) = \max(h(A,B), h(B,A)) \tag{12}$$

$$BCELoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{13}$$

where $h(A,B) = \max_{a\in A}\{\min_{b\in B}\|a-b\|\}$, $h(B,A) = \max_{b\in B}\{\min_{a\in A}\|b-a\|\}$ and $\|a-b\|$ is the one-way Hausdorff distance between set A and set B, and $\|b-a\|$ is the one-way Hausdorff distance between set B and set A.

3.4. Experimental results and analysis

To verify the performance advantages of the medical image segmentation model proposed in this paper, we compare it with several other classical medical image segmentation models, including U-Net, U-Net++, Attention U-Net, Ce-Net, UNeXt, and CMUNeXt. Table 2 provides the performance comparison of the different models on the Luna for Liver Images and Isic for Skin Damage2017 performance comparison on the dataset.

Table 2. Comparison of experimental data from different methods on the Luna and Isic2017 datasets.

Networks	Params (M)	Luna					Isic2017				
		Iou↑	Dice↑	hd↓	Loss↓	Time	Iou↑	Dice↑	hd↓	Loss↓	Time
U-Net	34.52	92.89	95.96	6.294	1.359	12.69	79.29	83.95	6.709	2.944	37.34
U-Net++	26.90	93.38	96.24	6.258	0.935	11.84	80.31	87.49	4.705	3.274	36.68
Attention-UNet	34.87	93.67	96.75	6.208	1.283	12.47	81.72	88.99	5.746	2.744	37.49
Ce-Net	18.63	93.64	96.72	6.209	0.531	11.08	81.05	88.63	5.708	1.835	32.25
UNeXt	**1.47**	94.26	96.80	6.132	0.801	**9.65**	82.66	88.85	4.733	1.767	**27.21**
CMUNeXt	3.14	94.91	97.03	6.178	0.822	10.71	83.05	89.01	4.686	1.046	28.09
DAU-Net	3.50	**95.46**	**97.36**	**5.874**	**0.463**	10.28	**83.36**	**89.22**	**4.657**	**0.730**	29.42

*Note: Time is the network processing time in h, Params(M) is the number of learnable parameters of the network.

From the experimental results in Table 2, it can be seen that the DAU-Net model proposed in this paper has better results in the four metrics of Iou, Dice, Hd, and loss on the Luna and Isic2017 datasets compared to the other models. However, in terms of network learnable parameters, the UNeXt network has fewer parameters and therefore has a shorter runtime. In the base network U-Net, the convolution operation used is a local operation, which does not have a larger sensory field and cannot fully tap into enough contextual information, so there is a loss of semantic information, and in the subsequent improvement of the network, the effect has increased significantly. In contrast, our model shows better performance in the Luna dataset. DAU-Net improves the Iou and Dice by 0.5 and 0.33%, respectively, compared to the CMUNeXt network, while reducing the loss by 0.359. For the Isic2017 dataset, our proposed model also has a better performance, but since the Isic2017 dataset is large, the running time is relatively long. Figure 3 shows some segmentation results.

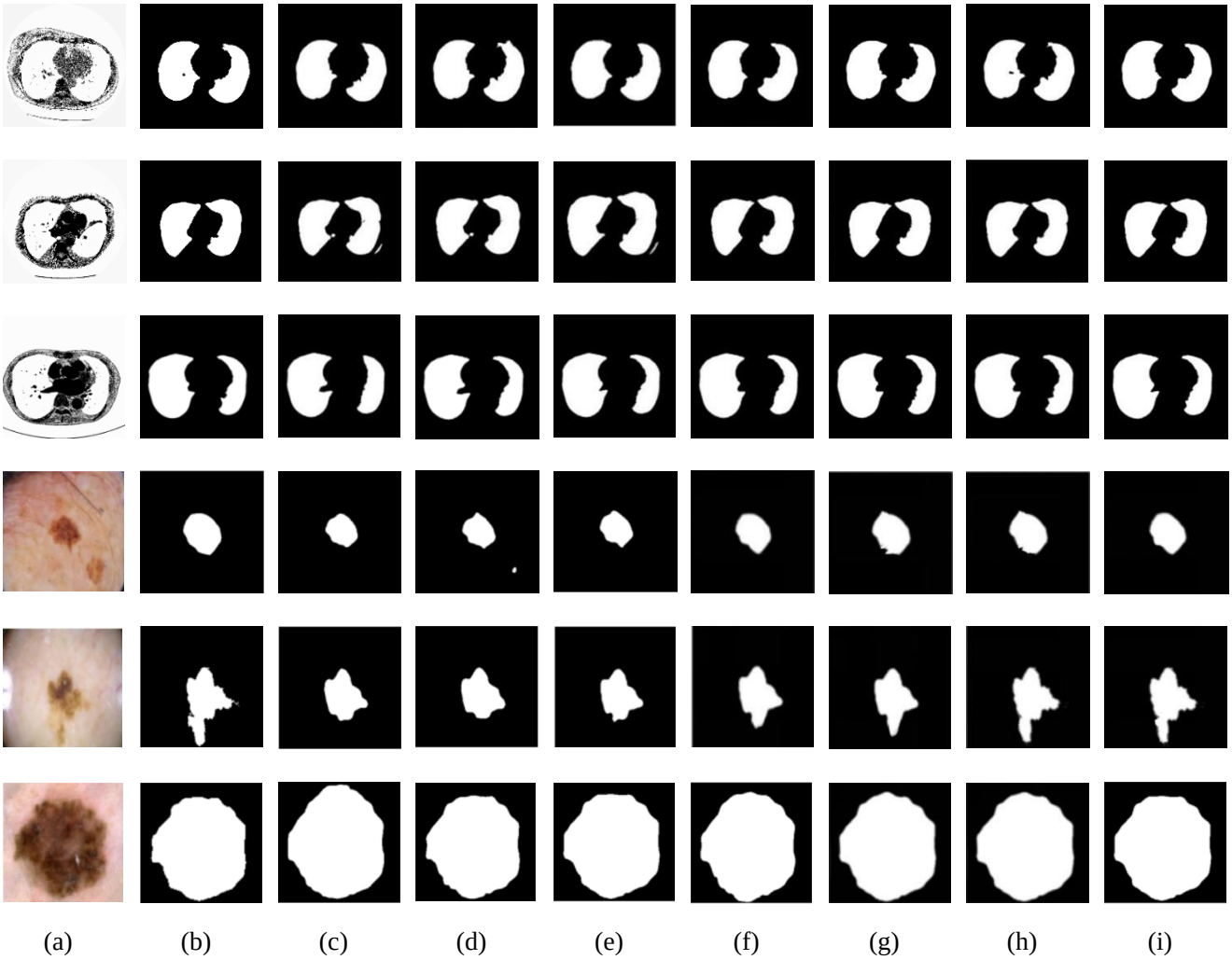

Figure 3. Segmentation results of different models on Luna and Isic2017 datasets, where (a) original graph; (b) Ground Truth; (c) U-Net; (d) U-Net++; (e) Attention-UNet; (f) Ce-Net; (g) UNeXt; (h) CMUNeXt; (i)DAU-Net.

3.5. Ablation experiments

Table 3. Results of ablation experiments.

Model	Iou↑	Dice↑	hd↓	Loss↓
U-Net	0.9289	0.9596	6.2947	1.3591
U-Net+stem	0.9357	0.9643	6.2140	1.3095
U-Net+hp	0.9346	0.9625	6.2142	1.3142
U-Net+DSAG	0.9399	0.9683	6.0479	1.2947
U-Net+hp+DSAG	0.9468	0.9698	6.1943	1.0439
U-Net+stem+DSAG	0.9519	0.9716	5.9837	0.8046
U-Net+stem+hp+DSAG	**0.9546**	**0.9736**	**5.8735**	**0.4634**

*Note: stem denotes encoder head module; hp denotes The Hadamard product; DSAG denotes Dual Scale Attention Gate.

In order to verify the validity of multiple modules of the experimental model in this paper, ablation

experiments are conducted on the stem, the Hadamard product, and dual-scale attention gate modules based on the U-Net network. The ablation experiments are conducted on the Luna dataset and the results are shown in Table 3.

From the data of the ablation experiments, it can be seen that when only stem and hp are added to the U-Net network, the performance of the network is only improved in a relatively small way, which is due to the fact that the extracted feature information is not fully utilized in the enhancement of the network's feature extraction, and so the performance of the network is not particularly improved. When the DSAG module is added, the network enhances the extracted information with features while expanding the sensory field to filter out the more valuable features, and it can be seen that the model improves by 1.1% in Iou value compared to the U-Net, indicating that the module plays a certain positive influence in the network. When all the modules are added together, stem reduces the loss of input image feature information, and hp generates feature representations with more representational ability, which is used to improve the understanding of different levels and semantic information, and finally, when going through DSAG, features are fused with contextual information from different scales, and the performance of the attention mechanism is exerted to enhance the target features after enough features are extracted, and from the experimental result From the experimental results, the Iou and Dice coefficients are improved by 2.57 and 1.4%, respectively, compared to U-Net network, which shows the better segmentation effect and better performance of our proposed model.

4. Discussion

In order to validate the performance of our proposed network, we have extensively studied the existing SOTA methods. At the same time we also scrutinized the PSNet network, which we know from the literature [26] is based on a dual-encoder-decoder architecture where the attention mechanism is incorporated into almost every module and level of the network through skip connections between the PS codecs, which allows the network to better capture the important features in the image. The difference with the model proposed in this paper is that DAU-Net utilizes a dual-scale attention gate, which improves the attention to global and local structures through a combination of point-by-point convolution and dilation convolution. From the experimental results, although the model shows better performance, it needs further enhancement and improvement in terms of edges and boundaries of the image, whereas PSNet uses the Transformer decoder as well as skipping the attention mechanism in the connections to enhance the attention to the critical regions. The model mainly focuses on the task of deep polyp segmentation of colonoscopy images, which is capable of identifying the location of polyps at the pixel level and distinguishing them from healthy tissues, and is suitable for the field of endoscopy. PSNet outperforms the current state-of-the-art results by comparing the existing five publicly available polyp datasets in an extensive study with performances of 0.863 and 0.797 in terms of mDice and mIoU, respectively. These two approaches, although different in their target domains, do share a common focus on key issues in image segmentation, which also inspires us to continue our research on image segmentation problems in the future.

A series of networks with superior performance have also emerged in the field of damage segmentation, including but not limited to AttentionU-net, CrackSegNet, Deeplab V3+, FPHBN, and U-Net. Among them, AttentionU-net, DeeplabV3+ and U-Net are the comparative experimental models in this paper. DeeplabV3+ [27] is a deep learning model for semantic image segmentation with a DCNN with null convolution and then a spatial pyramid pooling module with null convolution as its

main body, mainly to introduce multiscale information to enable the model to capture context from a variety of sensory fields. CrackSegNet [28] is a concrete crack segmentation network based on convolutional neural networks, which includes a backbone network, spatial pyramid pooling, and jump connection module, etc. FPHBN [29] is also a deep learning-based road crack detection method, which mainly integrates contextual information into low-level features in a feature pyramid fashion for crack detection. Although CrackSegNet and FPHBN are used for automatic crack segmentation tasks, there are some similarities with medical image segmentation in their network structures and feature extraction strategies. We can gain insight into their generalization and flexibility in different domains and consider whether they can also be useful in medical image segmentation tasks.

At the same time, we strongly agree that advanced networks based on attention modules have developed lightweight architectures with state-of-the-art performance. For example, [30] presents a network used to train an internal damage segmentation network (IDSNet). The network focuses on active thermography and uses an attention-based generative adversarial network (AGAN) to generate synthetic images. IDSNet exhibits a very high real-time processing capability, and its lightweight architecture results in a total number of learnable parameters of only 0.085 M. The network achieves significant performance gains on the internal damage segmentation task. [31] proposes a novel Semantic Translator Representation Network (STRNet) focusing on real-time crack segmentation at the pixel level. The network mainly consists of a squeezing and excitation attention-based encoder and an attention-based multi-head attention decoder, and the network achieves efficient and accurate crack segmentation while maintaining a fast processing speed. In addition, the network achieves the fastest processing speed of 49.2 frames per second by using a combination of techniques such as lightweight structures and optimized loss functions in its design, and STRNet shows the best performance in the evaluation metrics compared to other state-of-the-art networks. The discussion and study of these lightweight architecture networks with state-of-the-art performance also gives us important ideas for our subsequent research work.

In our future research, we will compare in detail the similarities and differences between our work and these state-of-the-art models; especially in terms of lightweight architecture, performance effects, and applicable scenarios. We will endeavor to ensure a thorough discussion of these related works in order to design networks with better performance.

5. Conclusions

In this paper, we propose the DSAG module to combine point-by-point convolution and dilation convolution to increase the receptive field without enlarging the feature dimensions, so that it captures more contextual information and realizes the fusion of feature mapping to improve the representational ability of the features. In addition, stem block and Hadamard product algorithms are introduced in the codec section in combination with convolution to avoid losing more feature information and at the same time generating more representational feature representations, which helps to deal with image variations at different scales. Based on this, we propose DAU-Net and show through experiments that our model can efficiently extract feature information, more accurately localize and segment the structure and lesion regions in medical images, showing better performance, and also can be applied to different scenarios. However, despite the success of our model, from the results, there is still room for improvement in edge and boundary detection; especially for some subtle structures in the edge region of the image, the model is fuzzy or inaccurate when performing segmentation. Therefore, in the

subsequent research, we will further optimize the DAU-Net model for edges and boundaries at different scales, so as to make it have a certain degree of continuity and stability, and to prevent the occurrence of broken or unnatural boundaries in the segmentation results. Eventually, the optimization will make it adaptable to more medical segmentation tasks.

Use of AI tools declaration

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

Acknowledgments

This work was partly supported by the Natural Science Basis Research Plan in Shaanxi Province of China under Grant 2023-JC-YB-517, and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems Beihang University under Grant VRLAB2023B08.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Wang, B. Ji, G. He, W. Yu, A computed tomography image segmentation algorithm for improving the diagnostic accuracy of rectal cancer based on U-net and residual block, *J. Biomed. Eng*., **39** (2022), 166–174. https://doi.org/10.7507/1001-5515.201910027
2. M. Yue, Q. Wei, W. Deng, T. Wang, Y. Deng, B. Huang, A review of automatic liver tumor segmentation based on computed tomography, *J. Biomed. Eng*., **35** (2018), 481–487. https://doi.org/10.7507/1001-5515.201708009
3. G. Dai, H. Sun, O. Yang, Research on image resolution problem in ultrasonic imaging d etection, *Comput. Knowl. Technol*., **6** (2010), 5937–5939. https://doi.org/10.3969/j.issn.1009 -3044.2010.21.115
4. W. M. Salama, A. Shokry, A novel framework for brain tumor detection based on convolutional variational generative models, *Multim. Tools Appl.*, **81** (2022), 16441–16454. https://doi.org/10.1007/s11042-022-12362-9
5. S. Sultana, A. Robinson, D. Y. Song, J. Lee, Automatic multi-organ segmentation in computed tomography images using hierarchical convolutional neural network, *J. Med. Imaging*, **7** (2020), 055001. https://doi.org/10.1117/1.JMI.7.5.055001
6. V. Venugopal, J. Joseph, M. V. Das, M. K. Nath, DTP-Net: A convolutional neural network model to predict threshold for localizing the lesions on dermatological macro-images, *Comput. Biol. Med*., **148** (2022), 105852. https://doi.org/10.1016/j.compbiomed.2022.105852
7. J. H. Pujar, P. S. Gurjal, K. S. Kunnur, Medical image segmentation based on vigorous smoothing and edge detection ideology, *Int. J. Electr. Comput. Eng.*, **4** (2020), 1143–1149.

8. K. Bhargavi, S. Jyothi, A survey on threshold based segmentation technique in image processing, *Int. J. Innov. Res. Dev.*, **3** (2014), 234–239.
9. T. A. Jemimma, Y. J. Vetharaj, Watershed algorithm based DAPP features for brain tumor segmentation and classification, in *2018 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, (2018), 155–158. https://doi.org/10.1109/ICSSIT.2018.8748436
10. S. Madhukumar, N. Santhiyakumari, Evaluation of k-Means and fuzzy C-means segmentation on MR images of brain, *Egyptian J. Radiol. Nuclear Med.*, **46** (2015), 475–479. https://doi.org/10.1016/j.ejrnm.2015.02.008
11. Z. Zheng, X. Zhang, S. Zheng, Y. Shi, CT liver image segmentation based on region growing and uniformized level set, *J. Zhejiang Univ.*, **52** (2018), 2382–2396.
12. B. Qian, Z. Xiao, W. Song, Improved convolutional neural network for segmentation on lung images, *Comput. Sci. Explor.*, **14** (2020), 1358–1367.
13. H. Li, Q. Zheng, W. Yan, R. Tao, X. Qi, Z. Wen, Image super-resolution reconstruction for secure data transmission in Internet of Things environment, *Math. Biosci. Eng.*, **18** (2021), 6652–6671. https://doi.org/10.3934/mbe.2021330
14. A. Rehman, M. Harouni, F. Zogh, T. Saba, M. Karimi, G. Jeon, Detection of Lung Tumors in CT Scan Images using Convolutional Neural Networks, in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023. https://doi.org/10.1109/TCBB.2023.3315303
15. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
16. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI* 2015, (2015), 234–241.
17. O. Oktay, J. Schlemper, L. Folgoc L, M. Lee, M. Heinrich, K. Misawa, et al., Attention u-net: Learning where to look for the pancreas, preprint, arXiv: 1804.03999.
18. Z. Zhou, M M R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, *IEEE Trans. Med. Imaging*, **39** (2019), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609
19. H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, et al., UNet 3+: a full-scale connected unet for medical image segmentation, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2020), 1055–1059. https://doi.org/10.1109/ICASSP40776.2020.9053405
20. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), 770–778.
21. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, V. K. Asari, Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, preprint, arXiv: 1802.06955.
22. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., Transunet: Transformers make strong encoders for medical image segmentation, preprint, arXiv: 2102.04306.
23. J. Ruan, M. Xie, J. Gao, T. Liu, Y. Fu, EGE-UNet: an Efficient Group Enhanced UNet for skin lesion segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2023), 481–490. https://doi.org/10.1007/978-3-031-43901-8_46

24. J. Valanarasu, M. Patel, Unext: Mlp-based rapid medical image segmentation network, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2022), 23–33. https://doi.org/10.1007/978-3-031-16443-9_3
25. F. Tang, J. Ding, L. Wang, C. Ning, S. K. Zhou, CMUNeXt: An efficient medical image segmentation network based on large kernel and skip fusion, preprint, arXiv: 2308.01239.
26. J. Lewis, Y. J. Cha, J. Kim, Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images, *Sci. Rep.*, **1183** (2023). https://doi.org/10.1038/s41598-023-28530-2
27. C. Liang-Chieh, Z. Yukun, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with atrous separable convolution for semantic image segmentation, in *Proceedings of the European conference on computer vision (ECCV)*, 2018.
28. R. Yupeng, H. Jisheng, H. Zhiyou, W. Lu, J. Yin, L. Zou, et al., Image-based concrete crack detection in tunnels using deep fully convolutional networks, *Constr. Building Mater.*, **234** (2020), 117367. https://doi.org/10.1016/j.conbuildmat.2019.117367
29. F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, H. Ling, Feature pyramid and hierarchical boosting network for pavement crack detection, *IEEE Trans. Intell. Transport. Syst.*, **21** (2020), 1525–1535. https://doi.org/10.1109/TITS.2019.2910595
30. R. Ali, Y. Cha, Attention-based generative adversarial network with internal damage segmentation using thermography, *Autom. Constr.*, **141** (2022), 104412. https://doi.org/10.1016/j.autcon.2022.104412
31. D. Kang, Y. Cha, Efficient attention-based deep encoder and decoder for automatic crack segmentation, *Struct. Health Monitor.*, **21** (2022), 2190–2205. https://doi.org/10.1177/14759217211053776