scientific reports

Check for updates

OPEN

A novel framework for segmentation of small targets in medical images

Longxuan Zhao[1,2]✉, Tao Wang[1,2], Yuanbin Chen[1,2], Xinlin Zhang[1,2,3], Hui Tang[1,2], Fuxin Lin[4,5,6,7,8], Chunwang Li[4,5], Qixuan Li[4,5], Tao Tan[9], Dezhi Kang[4,5,6,7,8]✉ & Tong Tong[1,2,3]✉

Medical image segmentation represents a pivotal and intricate procedure in the domain of medical image processing and analysis. With the progression of artificial intelligence in recent years, the utilization of deep learning techniques for medical image segmentation has witnessed escalating popularity. Nevertheless, the intricate nature of medical image poses challenges on the segmentation of diminutive targets is still in its early stages. Current networks encounter difficulties in addressing the segmentation of exceedingly small targets, especially when the number of training samples is limited. To overcome this constraint, we have implemented a proficient strategy to enhance lesion images containing small targets and constrained samples. We introduce a segmentation framework termed STS-Net, specifically designed for small target segmentation. This framework leverages the established capacity of convolutional neural networks to acquire effective image representations. The proposed STS-Net network adopts a ResNeXt50-32x4d architecture as the encoder, integrating attention mechanisms during the encoding phase to amplify the feature representation capabilities of the network. We evaluated the proposed network on four publicly available datasets. Experimental results underscore the superiority of our approach in the domain of medical image segmentation, particularly for small target segmentation. The codes are available at https://github.com/zlxokok/STSNet.

Medical image segmentation is a critical tool in computer-aided diagnosis. It aids in detecting and segmenting areas of pathology within images, facilitating the rapid identification of potential lesion regions. This has the potential to expedite diagnosis, increase the likelihood of lesion detection, reduce errors by clinical practitioners, and allow for more efficient use of their time in diagnostic assistance[1], benefiting both patients and doctors. Traditional medical image segmentation algorithms construct symmetrical bottom-up encoder-decoder structures to process images. These algorithms first compress input images into a latent space using encoders, then utilize decoders to learn positions of interest within the images. By introducing lateral signal propagation to vertical information flow, U-Net[2] architecture emerged as a seminal advancement in recent years. Today, most segmentation systems either include U-Net[2] or one of its variants, such as UNet++[3], 3D-UNet[4], UNeXt[5], HTC-Net[6], among others. O-Net[7] has shown excellent performance in tasks such as lesion segmentation[8] and pathology slice segmentation[9] and other organ segmentation[10]. A key aspect of the U-Net's success is its puzzlingly parameter-free nature. U-Net[2] doesn't estimate any non-convolutional trainable parameters within its architecture; it solely deals with convolutional parameters. U-Net[2] model, based on convolutional neural networks, has achieved significant breakthroughs in the accuracy and performance of medical image segmentation.

[1]College of Physics and Information Engineering, Fuzhou University, Fuzhou 350100, China. [2]Fujian Key Lab of Medical Instrumentation and Pharmaceutical Technology, Fuzhou 350100, China. [3]Imperial Vision Technology, Fuzhou 350100, China. [4]Department of Neurosurgery, Neurosurgery Research Institute, The First Affiliated Hospital, Fujian Medical University, Fuzhou 350100, China. [5]Department of Neurosurgery, National Regional Medical Center, Binhai Campus of the First Affiliated Hospital, Fujian Medical University, Fuzhou 350100, China. [6]Department of Neurosurgery, Fujian Institute of Brain Disorders and Brain Science, Fujian Clinical Research Center for Neurological Diseases, The First Affiliated Hospital and Neurosurgery Research Institute, Fujian Medical University, Fuzhou 350100, China. [7]Fujian Provincial Clinical Research Center for Neurological Diseases, The First Affiliated Hospital, Fujian Medical University, Fuzhou 350100, China. [8]Clinical Research and Translation Center, The First Affiliated Hospital, Fujian Medical University, Fuzhou 350100, China. [9]Macao Polytechnic University, Macao 999078, China. ✉email: zhaoxuanlong254@gmail.com; kdz99988@vip.sina.com; ttraveltong@gmail.com

However, the direct use of U-Net may not result in good performance in segmenting small targets, due to the fact that the multiple downsampling process in the encoding phase of U-Net may fully loss the information of small targets. Furthermore, the information of narrow and thin structures tends to be lost as networks become deeper due to convolution and pooling processes. A common strategy is to proportionally upscale input images to enhance the resolution of small targets or generate high-resolution feature maps[11]. Nonetheless, for extremely small lesion regions, which can account for less than 0.1% of the organ area, satisfactory results are still challenging to achieve using this approach. For example, we tried to apply the nnU-Net[12] for coarse and fine segmentation on the SCCM-2022 dataset in 3D volume. However, we found that in many cases, the coarse segmentation fails to identify the lesion area due to the small lesion pixel size on this dataset, making fine segmentation impractical. Although the direct use of nnU-Net for segmentation does not yield satisfactory results. Another promising solution is to introduce new network variants. DeepLabv3[13] has achieved excellent segmentation performance by incorporating dilated convolutions and multi-scale attention mechanisms. However, when applied to small target segmentation, the segmentation results of DeepLabv3 are inaccurate. Additionally, the small size of these objects might cause the loss of crucial information during feature extraction, consequently affecting segmentation accuracy and boundary clarity. Post-processing strategies like Markov random fields and conditional random fields have been proposed for improving the performance of small target segmentation[14]. However, since post-processing is not part of the segmentation network training, but rather applied to data after network output, the network cannot adapt its weights based on post-processing[15].

In order to address issues such as imbalanced samples of lesions and the tiny size of lesion areas, this paper proposes a segmentation network framework designed for small target regions in medical images.. It initially processes data through strategic augmentation, generalizing images at different levels, and then employs attention mechanisms to enhance the feature representation capabilities of convolutional neural networks.

Our key contributions are as follows:

- We introduce a novel data augmentation module. Under the premise of almost no increase in computational cost, with automatically identifying and amplifying the central region, images with extremely small targets and limited samples are effectively enhanced and generalized, significantly improving the segmentation accuracy of small targets in medical images.
- We propose an attention mechanism suitable for small target segmentation to better capture the interrelationships between images and channels. This attention surpasses in handling extremely small targets and a limited number of samples in medical images.
- Experimental results on the SCCM-2022, BUID-S, ISIC2017 and Lungs CT-Scan datasets demonstrate that this method outperforms other state-of-the-art segmentation techniques.

Related work

Medical image segmentation

Traditional medical image segmentation methods are diverse, and one of the most classic and widely used techniques is the thresholding method based on grayscale histograms[16], which aims to automatically select the optimal threshold for the image by maximizing the between-class variance, thereby segmenting the image into different regions. However, this method has certain limitations and performs poorly on images with high noise or small background-foreground differences. To address these issues, researchers have proposed many improvement strategies[17–19]. Li et al.[20] solved the threshold selection problem in image segmentation by minimizing the cross-entropy between the image and its regions. Walaa Ali et al.[21] addressed the poor image quality in Otsu's between-class variance estimation by selecting and combining two different mean filters to estimate the mean. However, traditional medical image segmentation methods have significant flaws, such as being unable to handle complex anatomical structures and being overly reliant on manual intervention.

Fully Convolutional Networks[22] and U-Net[2] have been commonly used in medical image segmentation, which frames the task as a dense classification problem. Current segmentation approaches can be broadly classified into two main categories: strategy optimization and network design. A novel strategy was proposed to handle the imbalanced training problem by optimizing the loss functions of the network[23]. In addition extracting uncertain pixels caused by factors like image noise and edge blurriness from high-frequency regions with rich details and complex textures can be utilized to further improve the segmentation performance[24].

The other approach focuses on optimizing the design of segmentation networks by integrating advanced techniques such as deformable convolutions[25], pyramid pooling[26], dilated convolutions[27], multi-scale context[28], attention mechanisms[29], and others to refine feature information. DoubleU-Net[30] combined two variations of the U-shaped architecture for biomedical image segmentation tasks. FD-UNet[31] incorporates dense connectivity in both the contraction and expansion pathways of U-Net architecture. This design can prevent the learning of redundant features and enhances the flow of information. CoTrFuse network[32] incorporates EfficientNet and Swin Transformer as dual encoders, fusing features from both branches with a CNN Fusion module before skip connections. This hierarchical structure offers flexibility in learning features across various scales and exhibits linear computational complexity relative to image size. These advantages make it perform well in visual tasks like image classification and semantic segmentation. However, most segmentation network designs are tailored for specific tasks, and there are few networks specifically designed for small targets. Additionally, due to the scarcity of cases and the extremely small lesion volumes in the SCCM-2022 dataset, conventional attention mechanisms cannot accurately extract network features, resulting in redundant information and blurred segmentation edges. In response to these challenges, we have designed an attention mechanism specifically for small targets to enhance network performance.

Small target segmentation

Due to operations such as convolutions and pooling within the network, as the network's depth increases, small and thin information can be lost[33]. Currently, various methods have been developed for addressing small target segmentation[34–36]. For instance[37], proposes a projection strategy in which 3D features are projected onto three orthogonal 2D planes to capture contextual attention from different views, filtering out redundant feature information to reduce the loss of key information in 3D scanning of small and medium-sized lesions. A data augmentation method based on a statistical deformation model was proposed[38]. CaraNet[39] presents a Context Axial Reverse Attention Network, utilizing axial reverse attention operations to detect both global and local feature information. CentroidNetV2[40] proposes a novel hybrid convolutional neural network designed specifically for segmenting and counting numerous small and connected object instances. By combining cross-entropy loss and Euclidean distance loss in the loss function, high-quality object instance centroids and boundaries can be obtained. RSTN[41] proposed a recurrent saliency transformation network to associate coarse and fine segmentation with saliency transformation modules to optimize the overall model. However, most processing strategies are performed directly on 3D dataset, requiring a significant amount of computing resources, and the results are not necessarily satisfactory[38].

Different network variants have been utilized[3,27,42], for small target segmentation. However, due to the limited informative content provided by small targets, the segmentation process becomes very challenging. A possible solution is post-processing[43], but since post-processing is not included in the segmentation model training, the model's output cannot conform to the weight values applied during post-processing[15]. Modifying loss functions presents another approach for addressing small target segmentation[44], and its advantage lies in avoiding an increase in the computational cost of the segmentation model. However, in medical image segmentation, developing a complex loss function specifically for specific task is undoubtedly time-consuming and labor-intensive, and may not necessarily achieve the desired results.[45,46].

Therefore, we propose a new approach that calculates the range defined as small targets by statistically analyzing the total lesion volume of the entire dataset. Subsequently, we apply data augmentation techniques such as cropping, scale augmentation, flipping, central expansion, and interpolation to small targets within that range to generate multiple enhanced images and labels. This diversifies the dataset and enhances the model's generalization ability.

Methods

Overall architecture design

Our approach follows a similar framework to the typical automated brain tumor segmentation process, as shown in Fig. 1. However, notable modifications have been introduced at each stage of this work to enhance the final segmentation performance. First, we transform 3D MRI images with sizes ranging from [256, 256, 16] to [640, 640, 35] into 2.5D images by concatenating three consecutive slices together to create a three-channel image, similar to a typical RGB image, in order to preserve sequential information. By performing preprocessing steps such as cropping, scaling, flipping, label dilation, center cropping, and interpolation on 2.5D images, we augment the dataset and generate focused and enhanced images. Lastly, we employ the Small Target Segmentation network (STS-Net) model to perform the segmentation of cerebral cavernous malformation, achieving the segmentation of lesion areas and healthy areas within the brain cavernous malformation. The following summary provides detailed information on each step.

Amplification blocks

We propose an augmentation module designed for small lesion targets with limited sample sizes, as shown in Fig. 2. For the SCCM-2022 dataset, we initially transform 3D MRI images with varying dimensions, ranging from [256, 256, 16] to [640, 640, 35], into multiple [480, 480, 1] 2D single-channel images. To preserve the sequential information, three consecutive slices are concatenated to create three-channel images, resembling typical RGB images, referred to as 2.5D images. Since only a few slices in most cases contain lesions in the 2.5D

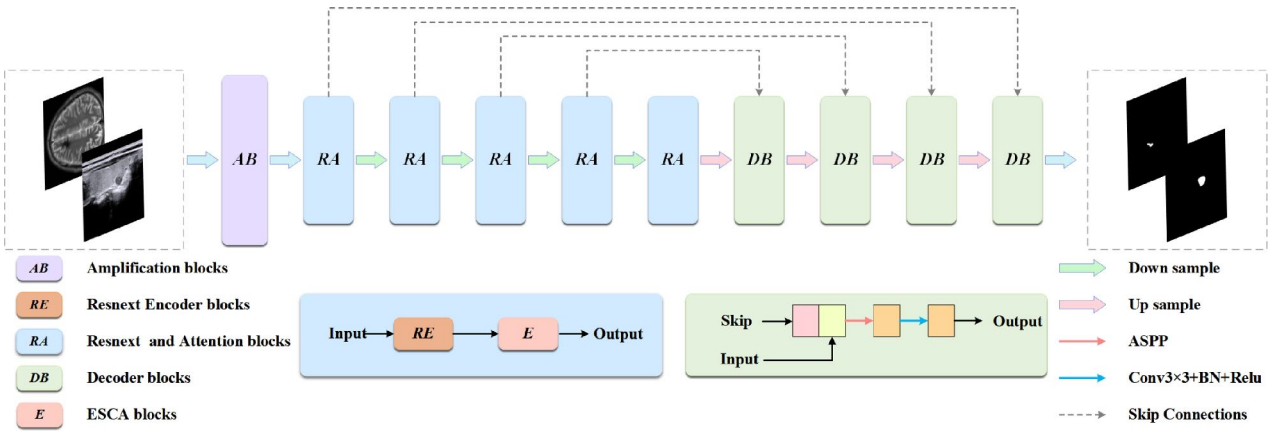

Fig. 1. Illustration of the proposed STS-Net.

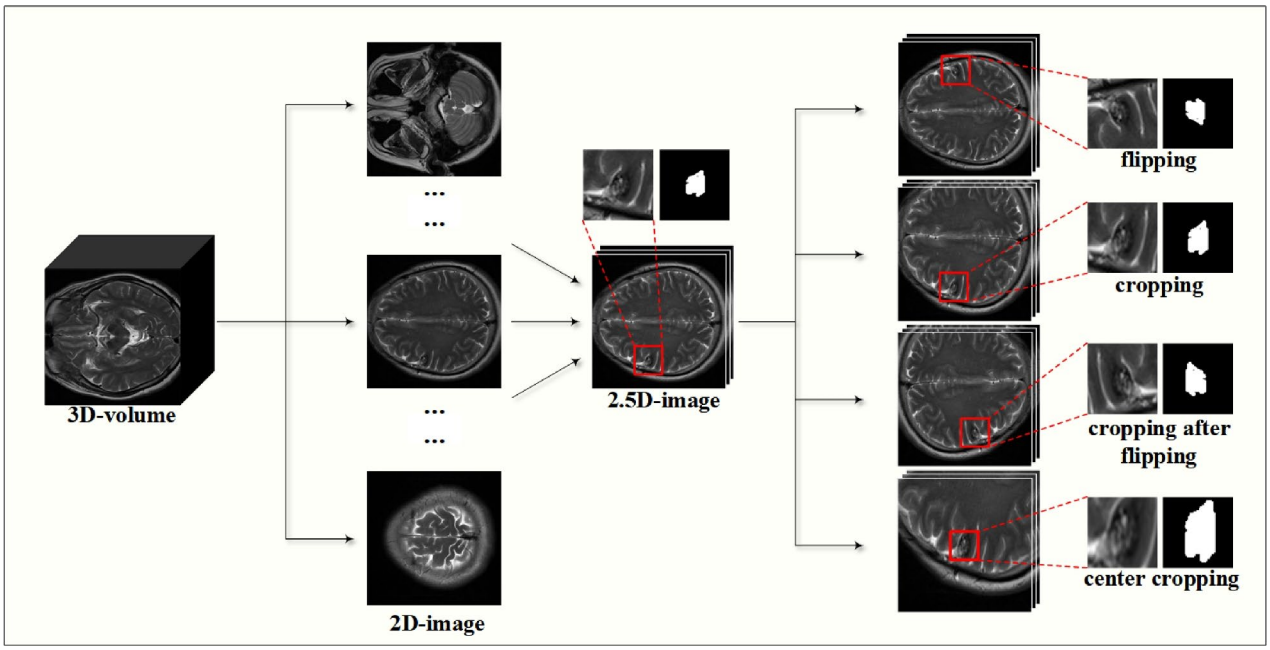

Fig. 2. Illustration of amplification blocks, 2.5D images are derived from three adjacent 2D slices, with the first images of the 2.5D images being the product of tripling the first slice of the 2D slice, and a similar process for the last images of the 2.5D images.

slices, we calculate the number of images containing lesion regions as A and the total area of lesions in 2.5D slices as S using the following equation:

$$N_1 = (\left\lceil \sqrt{\frac{S}{A}} \right\rceil)^2. \tag{1}$$

we can obtain a perfect square, denoted as N_1, where $\lceil \rceil$ denotes rounding up. When the lesion area in a 2.5D image is smaller than a certain value, we perform multiple augmentation operations on that image, as listed in Table 8. We employ various augmentation methods. Method one involves horizontal and vertical flipping, cropping, resizing, and flip cropping for images with lesion areas smaller than N_1. In the magnification process, the standard cropping operation first requires a boundary value, which was set to 50. Then, we crop 50 pixels from either the top and bottom edges, the left and right edges, or all four edges. The cropped image is then restored to its original size using linear interpolation, while the labels are restored using nearest-neighbor interpolation. In method two, By using the following equation:

$$N_2 = (\left\lceil \sqrt{\frac{N_1}{2}} \right\rceil)^2. \tag{2}$$

we can derive another perfect square number N_2. And applies similar augmentation techniques to smaller areas less than N_2. Method three involves additional cropping augmentation after applying the above augmentations to images with lesion areas smaller than N_2 to obtain finer-grained images. Method four focuses on center cropping in images with lesion areas smaller than N_2 extending the lesion area outward randomly in terms of dimensions before cropping and combining it with the images augmented by method three. The core advantage of this module lies in its ability to effectively enhance the diversity of medical image data, while preserving the finer details of small lesions.

ResNeXt blocks

Traditional methods for improving accuracy in models often involve increasing the depth or width of the network. However, as the number of hyperparameters increases (such as channel count and convolutional kernel size), the difficulty of network design and computational costs also rise. The ResNeXt[47] structure offers a solution to enhance accuracy without escalating parameter complexity, while also reducing the number of hyperparameters. As shown in Fig. 3, a ResNeXt block comprises a series of convolutional layers, pooling layers, batch normalization, softmax, and more. Each ResNeXt block introduces multiple branches (referred to as "cardinality") to enhance the feature learning capacity. These branches operate in parallel, each applying a set of transformations to the input feature map. The outputs of these branches are then aggregated through

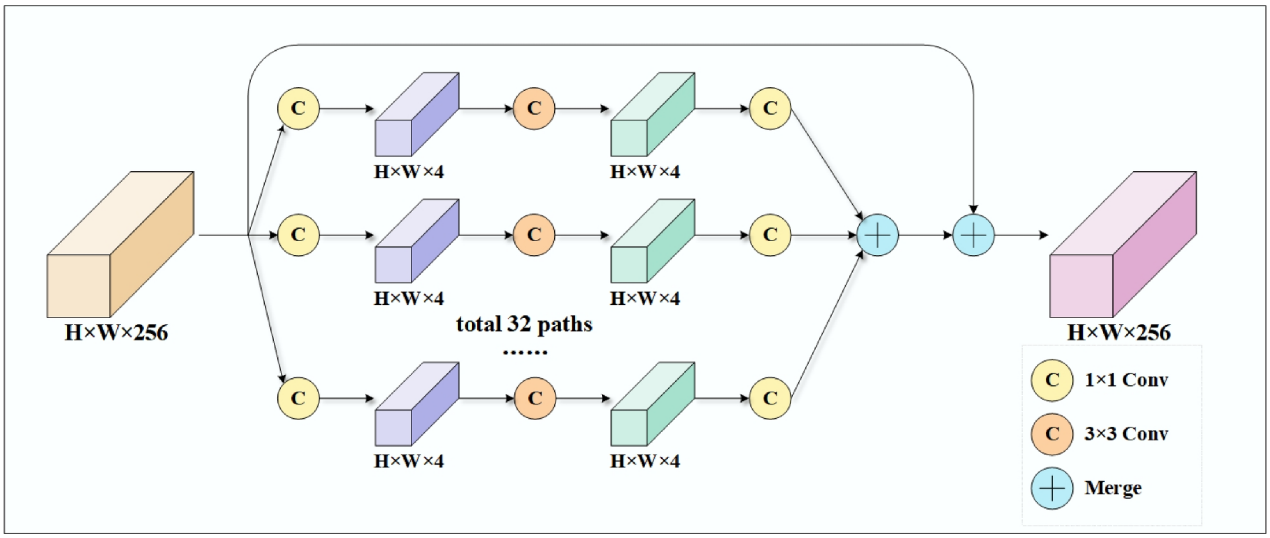

Fig. 3. Illustration of ResNeXt blocks.

summation, ensuring that the model can efficiently learn complex patterns. The output of a cardinality block can be expressed as:

$$F(x) = \sum_{i=1}^{C} \tau_i(x). \tag{3}$$

Here, C denotes the cardinality, which controls the number of parallel branches, and $\tau_i(x)$ represents the transformation applied by the i-th branch. The residual connection x ensures that the gradient can propagate more effectively during training, mitigating the vanishing gradient problem. These branches collaboratively learn features and pass the formula on the last layer:

$$y = x + \sum_{i=1}^{C} \tau_i(x). \tag{4}$$

This module addresses the challenges posed by increasing network depth while maintaining a tradeoff between parameter efficiency and computational complexity. Additionally, the modular design of ResNeXt allows for easy scalability and adaptation to different tasks, making it a versatile choice for various deep learning applications. Therefore, for the feature extraction module, we employ the ResNeXt block. By leveraging its multi-branch architecture and residual connections, our proposed system achieves superior feature representation while maintaining computational efficiency.

ESCA blocks

The ESCA (Efficient Small Channel Attention) layer functions as an attention mechanism are incorporated into convolutional neural networks, which can amplify the network's capacity to learn the comprehend relationships between channels. The incorporation of the ESCA layer refines channel attention mechanisms to effectively capture correlations among distinct channels in images, thereby enhancing the network's effectiveness across diverse computer vision tasks. As illustrated in Fig. 4, where the input features are denoted as f^i, the processed feature f_n can be derived through the following method:

$$f_1^i = GAP(f^i). \tag{5}$$

$$f_2^i = \frac{f_1^i + 2 \times P - D \times (K-1) - 1}{S} + 1. \tag{6}$$

$$f_3^i = \frac{f_2^i + 2 \times P - D \times (K-1) - 1}{S} + 1. \tag{7}$$

$$w_i = \frac{1}{1 + e^{-f_3^i}}. \tag{8}$$

$$f_n = f^i \times w_i. \tag{9}$$

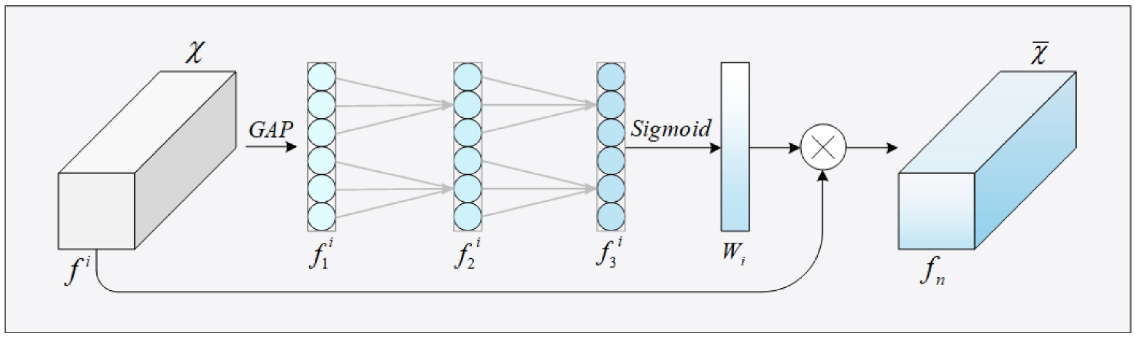

Fig. 4. Illustration of ESCA blocks, in GAP, the global average pooling operation is performed, and Sigmoid is a normalization function.

where *GAP* represent the global average pooling operation is performed, *P* represent the parameter filled outward, *K* represent the size of the convolution kernel, *D* represent the dilation of the convolution kernel, and *S* represent the step size. The use of GAP ensures that spatial information is aggregated into a compact representation, while the subsequent convolutional operations dynamically adjust the kernel size based on the number of channels, allowing the model to adaptively focus on relevant features. ESCA builds upon the SE module by substituting the fully connected (FC) layer, utilized in SE to learn channel attention information, with a 1×1 convolutional layer. This replacement not only reduces computational overhead but also enables the model to capture local channel dependencies more effectively. Initially, the input feature map f^i undergoes global average pooling, transforming the feature map from a [h, w, c] matrix to a [1, 1, c] vector f_1^i. This step ensures that the spatial dimensions are compressed, allowing the model to focus solely on channel-wise relationships. ESCA calculates an adaptive one-dimensional convolution kernel size based on the number of channels in the feature map. The determined kernel size is then applied in a one-dimensional convolution operation, and this operation is repeated twice, producing channel-wise weights f_3^i for each channel. The repetition of the convolution operation enhances the model's ability to capture complex channel interactions, while the use of a 1D convolution ensures computational efficiency. These channel-wise weights are subsequently normalized using the Sigmoid function, resulting in w_i. The Sigmoid function ensures that the weights are scaled between 0 and 1, allowing the model to selectively emphasize or suppress specific channels based on their importance. Finally, the normalized weights f^i are multiplied with the original input feature map channel-wise as w_i, generating a weighted feature map f_n. This weighted feature map emphasizes relevant channels while suppressing less important ones, thereby improving the overall discriminative power of the network.

During the convolution operation, the size of the kernel influences the receptive field. To tackle the challenge of extracting features from different ranges of input feature maps, ESCA employs dynamic convolution kernels for 1×1 convolutions. This dynamic approach learns the significance of different channels, ultimately enhancing the performance of tasks such as small target segmentation. By focusing on channel-wise relationships and employing adaptive mechanisms, ESCA significantly improves the network's ability to handle complex visual tasks, particularly those involving fine-grained details.

Experiments and results

Datasets

SCCM-2022 dataset

This dataset originates from the First Affiliated Hospital of Fujian Medical University and encompasses 113 3D-MRI images of individual cerebral cavernous malformations (CCM), We extracted the T2 sequence for experimentation. The image dimensions span from [256, 256, 16] to [640, 640, 35]. It encompasses 102 training images and 11 testing images. Each MRI image connects with manually annotated expert labels of cerebral cavernous malformations, designed for segmentation tasks. Within the complete 3D image, the largest lesion area constituted merely 0.03%, and the smallest comprised 0.0024%. In 2.5D images, the most substantial lesion area constituted 2.197%, and the smallest lesion area constituted only 0.017%.

BUID-S dataset

The second biomedical image dataset, known as the BUID-S dataset, comprises ultrasound images of all lesions less than 5.5% within the publicly accessible BUID dataset[48]. It incorporates 461 images, with 415 images designated for training and 46 images for testing. The BUID dataset is divided into three categories: benign, malignant images, and normal.

ISIC2017 dataset

The ISIC2017 dataset[49] contains 3 types of diseases: melanoma, seborrheic keratosis, and benign conditions, with a total of 2750 images. Among them, the training set consists of 2000 images, while the validation and test sets contain 150 and 600 images, respectively. Among them, the lesion area occupies more than 93% in the largest case, and only 0.3% in the smallest case. We normalized the image size to 512x512 and performed augmentation on images where the total number of white pixels is less than 2000.

Lungs CT-Scan dataset

Since the lesion regions in lung nodule segmentation tasks are small, we found a dataset for this task on Kaggle, called the Lungs CT-Scan dataset, and the dataset address is at https://www.kaggle.com/datasets/hasan1101/luna-classification-0. The dataset contains 2169 images, with 1,736 images used for training and 433 images for validation. Among the samples, the one with the largest lesion region has 4,862 white pixels, accounting for 1.854%, while the sample with the smallest lesion region has only 49 white pixels, representing just 0.02%. We normalized the image size to 512 × 512 and performed augmentation on images where the total number of white pixels is less than 576.

Evaluation metrics

To assess the segmentation performance of the model, we utilized several metrics, including the mean mDice coefficient (mDice), mean Intersection over Union (mIoU), Jaccard, Recall, and Precision. These metrics are associated with four values: True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN). The calculation methods for these metrics are as follows:

$$mDice = \frac{1}{k+1}\sum_{0}^{k}\frac{2\times TP}{2\times TP+FN+FP}. \tag{10}$$

$$mIoU = \frac{1}{k+1}\sum_{0}^{k}\frac{TP}{TP+FN+FP}. \tag{11}$$

$$Jaccard = \frac{TP}{TP+FN+FP}. \tag{12}$$

$$Recall = \frac{TP}{TP+FN}. \tag{13}$$

$$Precision = \frac{TP}{TP+FP}. \tag{14}$$

Implementation details

STS-Net was implemented using Python 3.8 and PyTorch 1.8.1. The training and testing platform ran on the Ubuntu 20.04 operating system and was equipped with an Nvidia RTX 3090 GPU with 24GB of memory. For the SCCM-2022 dataset, the input image size was configured in a size of 480× 480 pixels. In the case of the BUID-S dataset, the input image size was set to 512×512 pixels. During the training phase on the entire SCCM-2022 dataset, the model was optimized using the AdamW optimizer, with a learning rate and a batch size were set to 1e-4 and 4 respectively. The training phase consisted of 100 epochs. The training settings for the BUID-S, ISIC2017 and Lungs CT-Scan dataset were the same as those for the SCCM-2022 dataset, with the only difference being that the input image size was set to 512×512; ISIC2017 and Lungs CT-Scan dataset was also set to 512× 512. In terms of augmentation, smaller lesions in all four datasets were subjected to four standard augmentations and one center crop augmentation, meaning that each image meeting the augmentation criteria was expanded into six images for training.

Experimental results

We compared the proposed STS-Net model with several segmentation methods, including U-Net[2], DeepLabV3[13], UNet++[3], M^{2}SNet[45], MSGSE-Net[50], R2U-Net[51], DenseASPP[52], DeepLabV3AFMA[33], EH-former[53], UM-Net[54], I2U-Net[55], CaraNet[39] and CentroidNetV2[40]. We evaluated the segmentation performance of these methods using five metrics: mean Dice coefficient (mDice), mean Intersection over Union (mIoU), Jaccard, Recall, and Precision.

Results on the SCCM-2022 dataset

The segmentation results using different methods with or without augmentation on the SCCM-2022 dataset are compared in Table 1. It's evident that our STS-Net model achieved the best performance for the task of segmenting Cerebral Cavernous Malformation. The segmentation metrics achieved by our model on the SCCM-2022 dataset, in terms of mDice, mIoU, Jaccard, Recall, and Precision metrics are 82.13%, 77.45%, 55.12%, 62.96%, and 68.56%, respectively, outperforming all compared methods. Our proposed model exhibited the best overall performance across the five metrics. Additionally, our proposed augmentation strategy significantly improved evaluation metrics on other networks. For M^{2}SNet, the mDice score increased by 4.58%, and for U-Net, the mDice score increased by 4.64%. For larger lesions, all methods can detect and segment lesions, but the segmentation is not accurate enough. For example, in the first case of Fig. 5, both M^{2}SNet and UNet++ have overall shapes that differ from the lesion area, either under-segmenting or over-segmenting many pixels, while our shape is closest to the ground truth. In the second and third cases with smaller lesion areas, DenseASPP and DeepLabV3AFMA fail to detect the lesion area. UNet++ and DeepLabV3 show partial missed detections. In contrast, our method and M^{2}SNet not only successfully detect the lesions but also provides accurate segmentation results. This indicates that our method can accurately predict the location and boundaries of cerebral cavernous malformations, especially in small target segmentation.

In addition, we present the Params and FLOPs consumed by each model in the last two columns of Table 1. Although I2U-Net achieved suboptimal Params and optimal FLOPs, its segmentation performance is far

Models	mDice(%)↑	mIoU(%)↑	Jaccard(%)↑	Recall(%)↑	Precision(%)↑	Params(M)	FLOPs(G)
R2U-Net	64.85	61.91	24.22	30.58	35.65	39.09	2151.94
R2U-Net*	64.99	61.49	23.72	27.22	41.16		
MSGSE-Net	70.29	66.92	34.13	37.17	46.47	33.85	1273.24
MSGSE-Net*	72.58	68.98	38.16	42.40	55.65		
EH-former	72.33	68.88	38.03	44.56	49.71	86.37	275.52
EH-former*	73.05	68.38	37.08	42.44	59.67		
DenseASPP	72.78	68.63	37.56	43.91	51.04	46.15	221.66
DenseASPP*	73.61	69.67	39.60	45.78	53.24		
DeepLabV3AFMA	74.38	70.08	40.45	45.65	59.57	39.15	2413.72
DeepLabV3AFMA*	74.69	71.01	42.26	49.49	53.46		
CentroidNetV2	73.79	69.80	40.01	46.25	52.46	31.03	677.64
CentroidNetV2*	75.60	71.68	43.37	52.90	49.99		
UM-Net	73.45	69.71	39.65	46.59	51.39	22.78	144.76
UM-Net*	76.11	72.11	44.44	53.72	53.81		
I2U-Net	73.96	69.86	40.00	45.78	56.52	27.49	**98.43**
I2U-Net*	76.17	71.92	44.11	53.42	56.36		
CaraNet	75.03	70.94	41.69	48.82	53.71	46.64	162.17
CaraNet*	76.32	72.07	44.55	50.30	57.51		
U-Net	71.77	68.86	37.59	46.18	45.93	31.03	679.66
U-Net*	76.41	72.36	44.95	49.96	60.27		
DeepLabV3	75.56	71.56	43.38	50.61	55.44	**18.85**	724.37
DeepLabV3*	76.87	72.31	44.89	50.60	63.96		
UNet++	74.49	71.09	42.42	49.11	52.81	36.62	1945.86
UNet++*	77.67	73.46	47.15	53.17	62.15		
M^2SNet	74.89	70.82	41.92	52.04	51.96	29.74	248.83
M^2SNet*	79.47	74.69	49.63	54.93	**70.32**		
STS-Net(ours)	77.34	73.66	47.55	54.71	49.33	39.10	571.35
STS-Net(ours)*	**82.13**	**77.45**	**55.12**	**62.96**	68.56		

Table 1. Segmentation performance of different methods with or without the augmentation strategy on the SCCM-2022 dataset, with * denoting the use of an amplification blocks. Significant values are in bold.

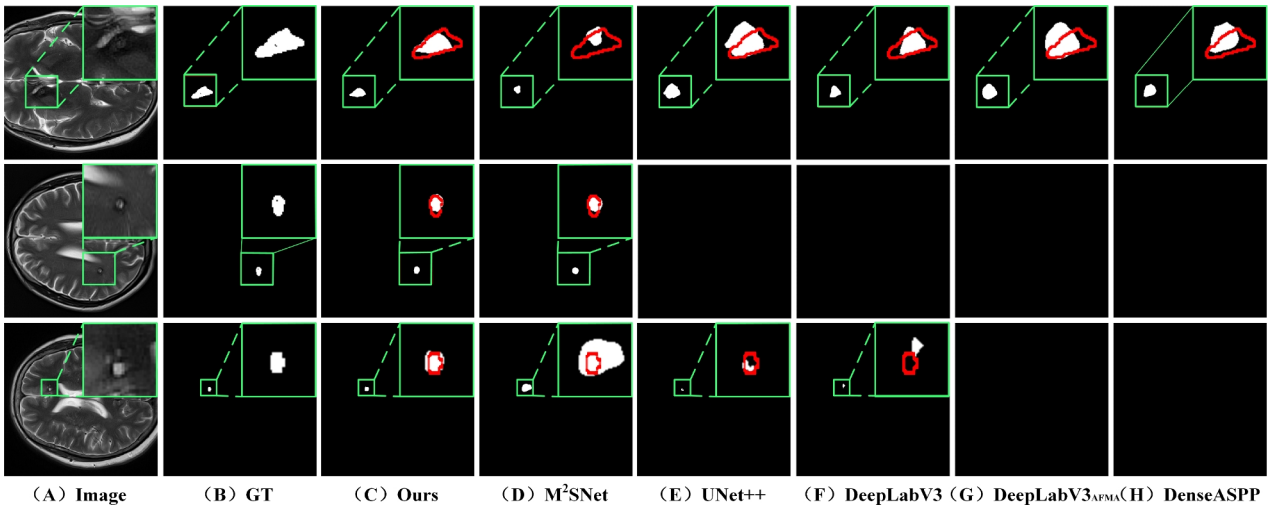

Fig. 5. Visual results of different methods on the SCCM-2022 dataset. The green-boxed area shown in the figure represents the region for visual zooming, the region outlined by the red lines depicts the actual values.

inferior to that of STS-Net, with a 5.96% lower mDice. STS-Net consumes more Params and is intermediate in FLOPs. This is because the ESCA attention mechanism, designed for small targets, requires multiple convolution operations, sacrificing some computational cost to improve model performance. Nevertheless, STS-Net can run on a single 8GB 2070 GPU and outperforms other models overall. This finding strongly highlights the ability of our STS-Net model to achieve excellent results while maintaining low computational complexity.

Result on the BUID-S dataset

Table 2 presents a comparison of the proposed STS-Net with previous state-of-the-art methods on the BUID-S dataset using both augmented and non-augmented data. The algorithm achieved good performance with metrics at 86.80% (mDice), 83.86% (mIoU), 68.57% (Jaccard), 72.56% (Recall), and 75.07% (Precision). Visualized segmentation results of different algorithms on this dataset are shown in Fig. 6, In the first case, all methods can accurately segment the lesion. Although our result is closest to the ground truth, the difference between our method and other methods are not significant. The M^2SNet and DeepLabV3 also perform well on this case. In the second case, M^2SNet, DenseASPP, DeepLabV3 and MSGSE-Net exhibit instances of under-segmentation. As indicated by the yellow boxes in the figure, DeepLabV3AFMA, M^2SNet, DeepLabV3, and MSGSE-Net have large areas of false positives. From the visual results, our STS-Net performs better in Breast ultrasound nodule segmentation, with segmentation results closer to the ground truth. This demonstrates the advantage of the proposed augmentation module in conjunction with the STS-Net architecture for small target segmentation.

Result on the ISIC2017 dataset

Due to the small sample size of the SCCM-2022 and BUID-S datasets, and the fact that most samples in the SCCM-2022 dataset are non-lesional, while the BUID-S dataset consists of small targets from the BUID dataset without sufficient large target samples, no validation set was created for either dataset. To fully evaluate the performance of STS-Net, we conducted experiments on the ISIC2017 dataset. The training set of the ISIC2017 dataset has a lesion area that exceeds 93% in the largest case, and only 0.3% in the smallest case, resulting in an imbalanced data distribution. We performed augmentation on samples with a lesion area smaller than 2000 after normalization, and the experimental results are shown in Table 3. Although STS-Net achieves the best performance on the ISIC2017 dataset, it does not outperform M^2SNet by much. This is primarily because the small target samples in the ISIC2017 dataset are too few, and STS-Net is a model specifically designed for small

Models	mDice(%)↑	mIoU(%)(%)↑	Jaccard(%)↑	Recall(%)↑	Precision(%)↑
R2U-Net	82.76	79.03	59.12	62.55	72.88
R2U-Net*	83.50	79.71	60.42	62.92	74.09
U-Net	83.42	80.04	61.10	67.75	72.21
U-Net*	83.84	80.41	61.76	68.08	73.06
UNet++	84.04	80.69	62.41	69.44	70.93
UNet++*	84.07	80.72	62.21	69.50	72.34
I2U-Net	83.53	80.16	61.33	65.46	72.35
I2U-Net*	84.41	81.07	63.16	69.05	72.24
MSGSE-Net	83.81	80.52	61.93	65.88	72.22
MSGSE-Net*	84.64	81.22	63.36	68.83	73.31
DeepLabV3	84.39	81.03	63.03	69.00	71.99
DeepLabV3*	85.00	81.50	63.94	69.35	73.74
UM-Net	83.17	79.52	59.76	64.90	71.79
UM-Net*	85.06	81.58	63.99	70.02	71.60
DenseASPP	84.70	81.53	64.08	68.43	72.32
DenseASPP*	85.24	81.79	63.87	68.21	75.18
CentroidNetV2	84.99	81.48	63.90	69.41	72.87
CentroidNetV2*	85.46	82.47	66.20	72.13	72.80
CaraNet	84.50	81.17	63.28	66.67	70.44
CaraNet*	85.66	82.69	66.43	71.11	73.00
EH-former	84.04	80.66	62.10	69.32	71.96
EH-former*	85.70	82.75	66.52	72.07	72.59
M^2SNet	85.69	82.77	66.47	**74.54**	71.35
M^2SNet*	86.22	83.13	67.17	72.55	73.98
DeepLabV3AFMA	78.33	75.22	51.20	57.19	63.26
DeepLabV3AFMA*	86.27	82.83	66.53	73.95	**76.52**
STS-Net(ours)	86.23	83.23	67.29	71.62	75.61
STS-Net(ours)*	**86.80**	**83.86**	**68.57**	72.56	75.92

Table 2. Segmentation performance of different methods and amplification or not on the BUID-S dataset, with * denoting the use of an amplification blocks. Significant values are in bold.

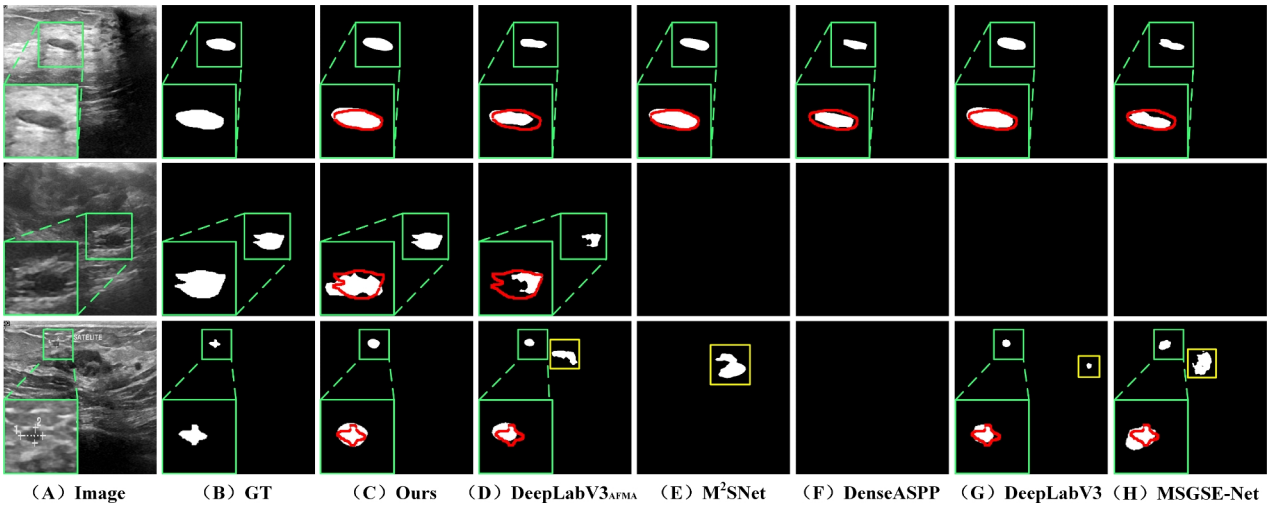

Fig. 6. Visual results of different methods on the BUID-S dataset. The green-boxed area shown in the figure represents the visual zoom-in region, the yellow-boxed area represents false positives, and the region outlined by the red lines depicts the ground truth values.

Models	mDice(%)↑	mIoU(%)(%)↑	Jaccard(%)↑	Recall(%)↑	Precision(%)↑
U-Net	81.23	73.54	63.66	77.98	81.19
U-Net*	82.61	75.29	64.56	73.31	87.08
UNet++	82.26	74.76	65.06	77.66	83.79
UNet++*	84.41	81.07	63.16	69.05	72.24
R2U-Net	82.95	75.58	64.56	72.00	89.22
R2U-Net*	84.89	77.68	57.79	76.28	88.59
MSGSE-Net	84.96	77.97	68.19	73.30	91.82
MSGSE-Net*	85.52	78.98	70.36	78.12	90.04
DeepLabV3AFMA	86.50	80.04	71.63	78.60	93.32
DeepLabV3AFMA*	87.71	81.46	72.89	77.66	92.99
EH-former	88.32	81.98	74.03	78.67	94.42
EH-former*	88.61	82.41	74.82	79.69	94.35
DenseASPP	88.31	82.02	74.37	79.44	93.91
DenseASPP*	88.76	82.51	75.18	80.30	93.99
CentroidNetV2	87.98	81.68	73.12	78.41	93.00
CentroidNetV2*	88.80	82.58	75.16	80.36	94.07
I2U-Net	88.67	82.41	74.78	78.88	95.13
I2U-Net*	88.87	82.68	75.16	79.75	94.63
DeepLabV3	87.97	81.65	73.58	78.87	92.61
DeepLabV3*	88.97	82.75	75.38	80.48	93.79
UM-Net	88.56	82.22	74.96	80.50	83.44
UM-Net*	89.05	82.88	75.66	79.54	**95.35**
CaraNet	88.38	82.16	74.26	79.93	91.91
CaraNet*	89.10	82.96	75.79	78.64	93.18
M^2SNet	88.82	82.57	75.07	79.05	94.59
M^2SNet*	89.35	83.22	76.15	81.03	94.03
STS-Net(ours)	89.42	83.34	76.38	82.77	92.63
STS-Net(ours) *	**89.69**	**83.70**	**76.84**	**82.24**	93.76

Table 3. Segmentation performance of different methods and amplification or not on the ISIC2017 dataset, with * denoting the use of an amplification blocks. Significant values are in bold.

target samples, so its performance on datasets with more large targets is inherently limited. The experimental results on the ISIC2017 dataset indicate that the augmentation module can balance the number of small and large target samples, thereby improving the model's performance to some extent. The experimental results are visualized in Fig. 7. In the first row of samples, other methods show varying degrees of false positives, while STS-Net demonstrates relatively accurate segmentation. In the second and third rows of samples, STS-Net achieves the best visualization results. The experimental results show that STS-Net and its augmentation module have certain advantages on datasets with more large targets.

Result on the Lungs CT-Scan dataset

Due to the extremely small lesion regions of lung nodules, the dataset is highly suitable for small-target image segmentation tasks, so we conducted experiments on the Lungs CT-Scan dataset. The lesion ranges in the Lungs CT-Scan dataset are similar to those in the SCCM-2022 dataset, but the sample distribution in the Lungs CT-Scan dataset is more balanced. We normalized and performed data augmentation on samples with lesion regions smaller than 576, and the experimental results are shown in Table 4. Due to the extremely small lesion regions, the loss functions of I2U-Net and UM-Net became excessively large in the first iteration of the training phase, preventing convergence and resulting in entirely black outputs on the validation set. The root cause is that, with equal loss function weights, the initial weight values for I2U-Net and UM-Net models were too large, leading to gradient explosion. In the segmentation experiments on the Lungs CT-Scan dataset, STS-Net still achieved the best performance, outperforming other models on four metrics. STS-Net without using augmented data already surpassed other models using augmented data, highlighting its significant advantage on small-target datasets. The visualization results of the experiments are shown in Fig. 8. For the small lesion samples in the first row, only STS-Net demonstrated relatively accurate segmentation, while other models exhibited clinically significant missed detections. For the lesion samples in the second and third rows, only STS-Net avoided clinically significant false detections, while other models showed substantial over-segmentation. The results confirm that STS-Net holds a certain advantage in small-target segmentation tasks.

Ablation study

An ablation study was performed to analyze the effectiveness and contribution of each module in STS-Net. The ablation study was performed on the SCCM-2022 dataset, and the results are presented in Tables 5, 6, 7, 8 and 9.

Effect of encoder in the STS-Net

The effect using different encoders was analyzed and the results are compared in Table 5. It can be seen that the ResNeXt block as the encoder framework among the five compared encoders. Upon comparing the accuracy of several encoders, including MobileNet-v2[56], DPN68[57], ResNet50[58], and EfficientNet-b0[59], it was discerned that ResNeXt50-32x4d[47] achieved the highest accuracy. Consequently, we employed the ReNeXt block as the encoder for STS-Net.

Effect of decoder in the STS-Net

When the ResNeXt50-32x4d was used as encoder, the results using different decoders are compared in Table 6. From the table, it can be seen that the accuracy of DeepLabV3, surpasses that of several other decoders such as DeepLabV3+[27], U-Net[2], PAN[60], and UNet++[3]. Based on this comparison, we chose DeepLabV3 as the decoder for STS-Net.

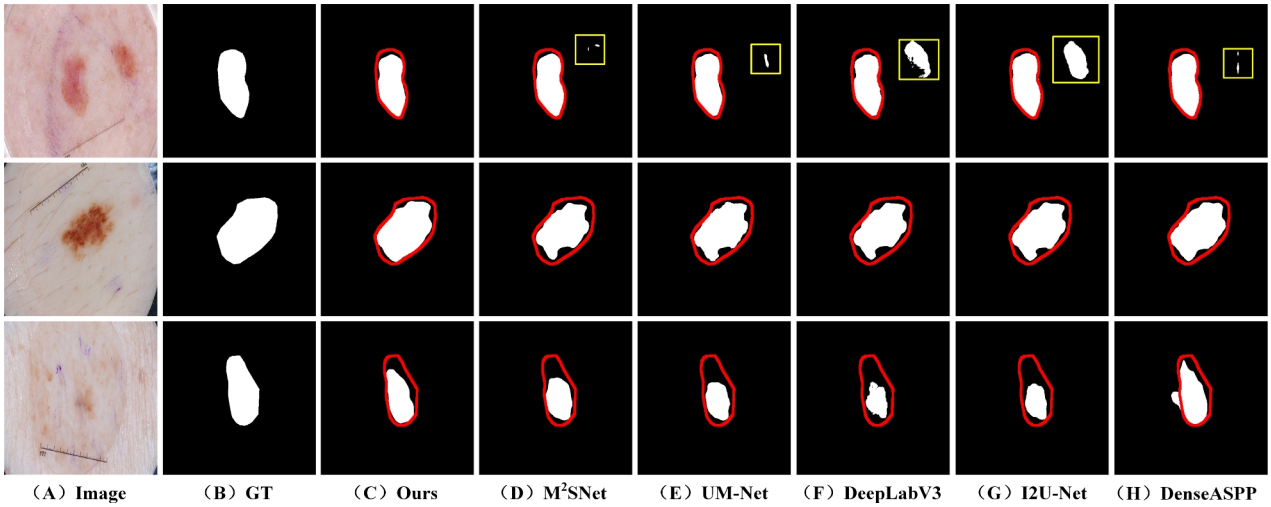

Fig. 7. Visual results of different methods on the ISIC2017 dataset. The yellow-boxed area represents false positives, and the region outlined by the red lines depicts the ground truth values.

Models	mDice(%)↑	mIoU(%)(%)↑	Jaccard(%)↑	Recall(%)↑	Precision(%)↑
I2U-Net	–	–	–	–	–
I2U-Net*	–	–	–	–	–
UM-Net	–	–	–	–	–
UM-Net*	–	–	–	–	–
R2U-Net	75.19	71.41	42.97	53.12	56.33
R2U-Net*	77.02	73.62	45.36	50.76	60.50
EH-former	83.42	79.26	58.61	66.40	71.94
EH-former*	84.05	79.89	59.87	69.04	71.89
DenseASPP	85.78	81.19	62.46	72.18	75.36
DenseASPP*	86.52	82.02	63.80	71.35	77.34
DeepLabV3AFMA	86.69	82.26	64.30	71.55	77.39
DeepLabV3AFMA*	87.30	83.01	66.96	74.00	77.70
MSGSE-Net	88.10	84.50	69.07	77.37	77.95
MSGSE-Net*	88.39	84.58	69.22	77.64	79.69
CentroidNetV2	87.75	84.50	69.07	77.37	77.95
CentroidNetV2*	88.45	84.75	69.60	78.25	80.03
U-Net	88.16	84.43	69.13	79.83	77.90
U-Net*	88.59	84.88	69.83	77.94	79.96
CaraNet	88.14	84.46	69.10	78.05	77.77
CaraNet*	88.61	84.90	69.86	78.36	81.20
DeepLabV3	88.24	84.58	70.23	77.61	80.65
DeepLabV3*	88.82	85.37	71.30	78.21	81.36
M^2SNet	88.24	84.67	70.41	74.74	**82.46**
M^2SNet*	88.78	85.70	72.47	78.01	81.13
UNet++	88.49	84.93	69.93	77.46	80.08
UNet++*	88.90	85.36	70.78	78.61	80.06
STS-Net(ours)	89.27	85.80	71.65	77.88	81.74
STS-Net(ours) *	**90.02**	**86.64**	**73.33**	**81.16**	81.31

Table 4. Segmentation performance of different methods and amplification or not on the Lungs CT-Scan dataset, with * denoting the use of an amplification blocks. Significant values are in bold.

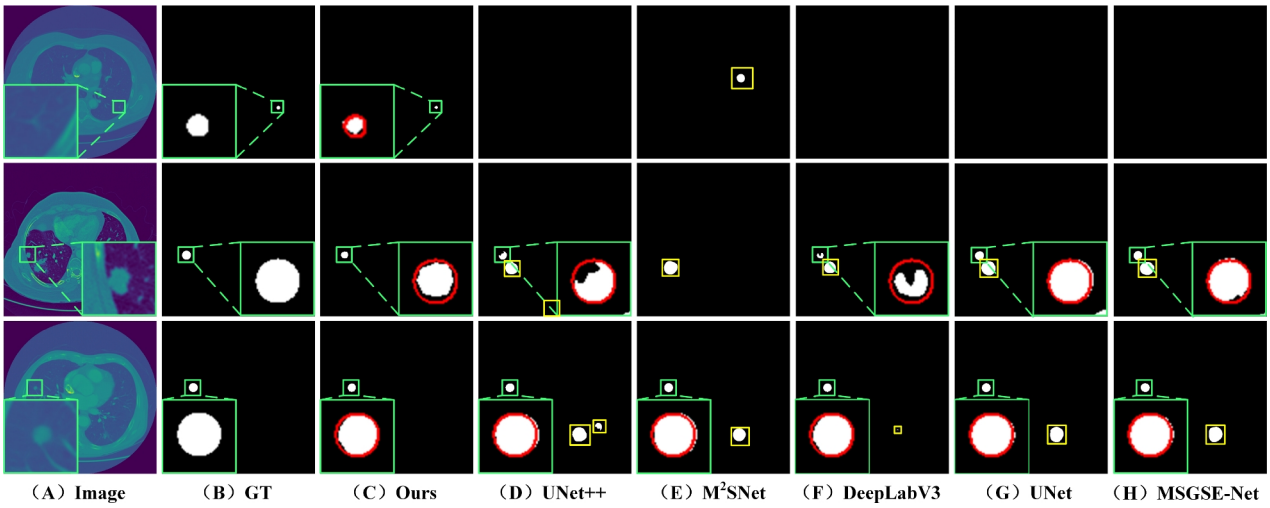

Fig. 8. Visual results of different methods on the Lungs CT-Scan dataset. The yellow-boxed area represents false positives, and the region outlined by the red lines depicts the ground truth values.

Effect of amplification methods in the STS-Net

When utilizing the ResNeXt50-32x4d encoder, DeepLabV3 decoder, and incorporating the ESCA module, we carried out an ablation study on data augmentation methods, and the results are presented in Table 7. Method 1 executes a series of operations, including horizontal and vertical flipping, cropping, resizing, and post-

Encoder	mDice(%)↑	mIoU(%)↑	Jaccard(%)↑	Recall(%)↑	Precision(%)↑
MobileNetV2[56]	65.80	62.06	24.47	26.67	42.21
EfficientNet-b0[59]	70.52	66.97	34.23	39.65	49.42
DPN68[57]	71.35	66.93	34.16	39.19	55.38
ResNet50[58]	75.74	71.98	44.20	51.03	56.10
ResNeXt50-32x4d[47]	**79.82**	**75.09**	**50.43**	**59.76**	**70.57**

Table 5. Ablation of STS-Net encoder based on SCCM-2022 dataset. Significant values are in bold.

Decoder	mDice(%)↑	mIoU(%)↑	Jaccard(%)↑	Recall(%)↑	Precision(%)↑
DeepLabV3+[27]	74.79	70.99	42.21	49.75	52.46
U-Net[2]	74.83	71.42	43.06	46.84	56.33
PAN[60]	75.12	71.24	42.70	47.33	57.28
UNet++[3]	75.34	71.05	42.34	46.10	60.67
DeepLabV3[13]	**82.13**	**77.45**	**55.12**	**62.96**	**68.56**

Table 6. Ablation of STS-Net decoder based on SCCM-2022 dataset. Significant values are in bold.

Augment	mDice(%)↑	mIoU(%)↑	Jaccard(%)↑	Recall(%)↑	Precision(%)↑
Baseline	77.34	73.66	47.55	54.71	59.33
Method 1	77.91	74.10	48.43	54.16	61.58
Method 2	79.33	75.06	50.35	58.90	64.37
Method 3	78.90	74.63	49.48	56.47	62.10
Method 4	**82.13**	**77.45**	**55.12**	**62.96**	**68.56**

Table 7. Ablation of STS-Net amplification methods based on SCCM-2022 dataset. Method 1 represents using normal amplification for images with illness areas less than 2025, method 2 means using normal amplification for images with illness areas less than 1024, method 3 means double quantity normal magnification was used for images with lesions smaller than 1024, method 4 means adding center amplification on method 2. Significant values are in bold.

Backbone	Amplification	ESCA	mDice(%)↑	mIoU(%)↑	Precision(%)↑
*			74.70	70.96	56.70
	*		75.84	71.52	61.99
		*	77.34	73.66	59.33
	*	*	**82.13**	**77.45**	**68.56**

Table 8. Ablation study for each component of STS-Net for Cerebral Cavernous Malformation segmentation on the SCCM-2022 dataset. Significant values are in bold.

Attention	mDice(%)↑	mIoU(%)↑	Jaccard(%)↑	Recall(%)↑	Precision(%)↑
CBAM[61]	70.00	66.32	32.88	36.48	48.45
SE[62]	75.03	70.82	41.91	46.69	62.06
None	75.84	71.52	42.50	46.52	61.99
GCT[63]	77.76	74.01	48.67	55.40	51.32
SRM[64]	78.77	74.53	49.05	56.10	61.43
ESCA	**82.13**	**77.45**	**55.12**	**62.96**	**68.56**

Table 9. Ablation of STS-Net attention mechanism based on SCCM-2022 dataset. Significant values are in bold.

flipping cropping, to magnify features for images where the lesion area falls below 2025 (45 × 45). Method 2 concentrates on even smaller regions, specifically those below 1024 (32 × 32) in size. It employs the same augmentation approach as Method 1 to boost image features. Method 3 extends Method 2 by applying the same augmentation and then further expanding the lesion area in images where it remains smaller than 1024 (32 × 32) after the initial augmentation, refining image features. Method 4 builds upon Method 3 by introducing central augmentation, where the lesion area of the image experiences cropping and then expands outward in random dimensions. Subsequently, this expanded area undergoes cropping, effectively amplifying the feature information of the lesion area. Experimental results on the SCCM-2022 dataset indicate that Method 4 produces superior enhancement results. The method we employed is Method 4, which utilizes central augmentation to magnify features in the lesion area.

Effect of each component in the STS-Net

Our proposed STS-Net model uses U-Net as the baseline model. Comparing the second and third rows in Table 8 with the first row, it can be observed that adding Amplification blocks or ESCA blocks to the encoder part of the network can improve segmentation performance. As shown in the fourth row of Table 8, compared to the baseline network (shown in the first row), the network with Amplification blocks and ESCA modules shows improvements in mDice, mIoU, and Precision. mDice increases from 74.70% to 82.13%, an improvement of 7.43%, mIoU increases from 70.96% to 77.45%, an improvement of 6.49%, and Precisiom increases from 56.70% to 68.56%, an improvement of 11.86%. These results indicate that the Amplification blocks and ESCA modules in our STS-Net are beneficial for small target medical image segmentation tasks, which indicates that these two modules can effectively learn global and local semantic information.

Effect of attention mechanism in the STS-Net

On the augmented SCCM-2022 dataset, we replaced the ESCA attention mechanism in STS-Net to highlight the advantages of ESCA attention, with the experimental results shown in Table 9. When using CBAM[61] and SE[62] attention, their performance was even lower than when no attention was applied, resulting in a negative growth of the model. When using GCT[63] and SRM[64] attention, although there was some performance improvement, they still did not outperform ESCA attention. The main reason is that we specifically designed the ESCA attention mechanism for small target datasets through multiple experiments, while other attention mechanisms were designed for their specific tasks and have lower adaptability to special tasks like SCCM-2022. The experimental results demonstrate the significant advantage of ESCA attention.

Discussion

In general, our amplification method employs center amplification and other techniques to magnify lesion feature information and balance the number of samples. Simultaneously, the ESCA attention mechanism can effectively recognize small target areas, augment the generalization capability of the medical image enhancement model, and consequently enhance segmentation accuracy. However, our method still has certain limitations. Our amplification blocks primarily addresses the issue of highly imbalanced sample sizes and extremely small lesion areas. However, when facing tasks such as expansive lesion areas, multiple lesion regions, and irregularly distributed vessel segmentation, our data amplification blocks performs poorly. For example, in the ISIC2017 dataset, some samples have lesion areas that occupy 80% of the total image area. In such cases, regular cropping may remove part of the lesion area, and center cropping is not applicable to images where the lesion area is adjacent to the image boundary. Moreover, in large target datasets, the formula computation for smaller targets in the dataset is not precise enough, and the ESCA attention mechanism does not yield significant benefits for large targets. STS-Net is also suitable for datasets where both large and small targets exist, but the number of small targets is limited. Our augmentation module can dynamically balance the sample size of the dataset to improve model performance. However, in datasets with fewer large targets, augmenting small targets will further imbalance the sample size, leading to a gradual decline in segmentation performance. Finally, the cerebral cavernous malformations cases we acquired are from a 3D dataset. Direct segmentation with other algorithms on 3D data does not perform well due to the small lesion areas, with some cases accounting for less than 0.1% of the total area. In such cases, simply predicting the model as all black results in an accuracy of 99.9%. Therefore, we process the dataset into 2D and then perform augmentation, reducing the computational load while making it easier to balance the sample size. In future, we will explore strategies and segmentation models that directly perform augmentation on 3D datasets, further improving the approach for small-target medical image segmentation tasks.

Conclusion

In this study, we proposed a novel approach, STS-Net, for small target medical image segmentation. Our method first passes through an effective Amplification blocks to enlarge the lesion area and enlarge the number of samples. It also incorporates the ESCA module in the encoding stage to obtain rich global and local semantic information, thus improving the network's segmentation performance. Experimental results on four different datasets, SCCM-2022, BUID-S ,ISIC2017 and Lungs CT-Scan, demonstrate that the proposed STS-Net method outperforms several state-of-the-art segmentation methods. The research findings emphasize the effectiveness of our proposed approach in medical image segmentation with extremely small target lesions and imbalanced sample sizes, offering assistance in diagnosing rare diseases with tiny lesions for medical professionals.

Data availability

STS-Net was evaluated using both internal and public datasets. The SCCM-2022 dataset that support the findings of this study are available from The First Affiliated Hospital, Fujian Medical University but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the author Zhao.L at 1154692412@qq.com. The BUID dataset is fetched from https://www.sciencedirect.com/science/article/pii/S2352340919312181 and processed by our code into the corresponding small target dataset BUID-S. ISIC2017 dataset is fetched from https://challenge.isic-archive.com/data/#2017. Lungs CT-Scan dataset address is fetched from https://www.kaggle.com/datasets/hasan1101/luna-classification-0.

Received: 8 June 2024; Accepted: 13 March 2025
Published online: 22 March 2025

References

1. Liang, D. et al. Coronary angiography video segmentation method for assisting cardiovascular disease interventional treatment. *BMC Med. Imaging* **20**, 1–8 (2020).
2. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. 234–241 (Springer, 2015).
3. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: A nested U-Net architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2018*. 3–11 (Springer, 2018).
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*. 424–432 (Springer, 2016).
5. Valanarasu, J. M. J. & Patel, V. M. UNeXt: MLP-based rapid medical image segmentation network. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*. 23–33 (Springer, 2022).
6. Tang, H. et al. HTC-Net: A hybrid CNN-transformer framework for medical image segmentation. *Biomed. Signal Process. Control* **88**, 105605. https://doi.org/10.1016/j.bspc.2023.105605 (2024).
7. Wang, T. et al. O-Net: A novel framework with deep fusion of CNN and transformer for simultaneous segmentation and classification. *Front. Neurosci.* **16**, 876065 (2022).
8. Xie, Y., Zhang, J., Xia, Y. & Shen, C. A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Trans. Med. Imaging* **39**, 2482–2493 (2020).
9. Hu, Z. et al. A multi-task deep learning framework for perineural invasion recognition in gastric cancer whole slide images. *Biomed. Signal Process. Control* **79**, 104261 (2023).
10. Zhou, X. et al. CUSS-Net: A cascaded unsupervised-based strategy and supervised network for biomedical image diagnosis and segmentation. *IEEE J. Biomed. Health Inform.* **27**, 2444–2455. https://doi.org/10.1109/JBHI.2023.3238726 (2023).
11. Li, J. et al. Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1222–1230 (2017).
12. Isensee, F. et al. nnU-Net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation. *arXiv preprint* arXiv:1809.10486 (2018).
13. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint* arXiv:1706.05587 (2017).
14. Chandra, S. & Kokkinos, I. Fast, Exact and Multi-scale inference for semantic image segmentation with deep Gaussian CRFs. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. 402–418 (Springer, 2016).
15. Guo, D., Zhu, L., Lu, Y., Yu, H. & Wang, S. Small object sensitive segmentation of urban street scene with spatial adjacency between object classes. *IEEE Trans. Image Process.* **28**, 2643–2653 (2018).
16. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66. https://doi.org/10.1109/TSMC.1979.4310076 (1979).
17. Jumiawi, W. A. H. & El-Zaart, A. Gumbel (EVI)-based minimum cross-entropy thresholding for the segmentation of images with skewed histograms. *Appl. Syst. Innov.* **6**, 87 (2023).
18. Kittaneh, O. A. The variance entropy multi-level thresholding method. *Multimed. Tools Appl.* **82**, 43075–43087 (2023).
19. Jumiawi, W. A. H. & El-Zaart, A. A boosted minimum cross entropy thresholding for medical images segmentation based on heterogeneous mean filters approaches. *J. Imaging* **8**, 43 (2022).
20. Li, C. & Lee, C. Minimum cross entropy thresholding. *Pattern Recognit.* **26**, 617–625. https://doi.org/10.1016/0031-3203(93)90115-D (1993).
21. Jumiawi, W. A. H. & El-Zaart, A. Otsu thresholding model using heterogeneous mean filters for precise images segmentation. In *2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE)*. 1–6. https://doi.org/10.1109/ICATEEE57445.2022.10093097 (2022).
22. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3431–3440 (2015).
23. Yadav, O. P., Bhamare, S. S. & Rathore, A. Reliability-based robust design optimization: A multi-objective framework using hybrid quality loss function. *Qual. Reliab. Eng. Int.* **26**, 27–41 (2010).
24. Lee, C. E., Park, H., Shin, Y.-G. & Chung, M. Voxel-wise adversarial semi-supervised learning for medical image segmentation. *Comput. Biol. Med.* **150**, 106152 (2022).
25. Yang, X., Li, Z., Guo, Y. & Zhou, D. DCU-net: A deformable convolutional neural network based on cascade U-net for retinal vessel segmentation. *Multimed. Tools Appl.* **81**, 15593–15607 (2022).
26. Huang, Z. et al. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. *Inf. Sci.* **522**, 241–258 (2020).
27. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*. 801–818 (2018).
28. Shi, Z. et al. MD-Net: A multi-scale dense network for retinal vessel segmentation. *Biomed. Signal Process. Control* **70**, 102977 (2021).
29. Liu, Y., Zhang, Z., Liu, X., Lei, W. & Xia, X. Deep learning based mineral image classification combined with visual attention mechanism. *IEEE Access* **9**, 98091–98109 (2021).
30. Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P. & Johansen, H. D. DoubleU-Net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. 558–564 (IEEE, 2020).

31. Guan, S., Khan, A. A., Sikdar, S. & Chitnis, P. V. Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal. *IEEE J. Biomed. Health Inform.* **24**, 568–576 (2019).
32. Chen, Y. et al. CoTrFuse: A novel framework by fusing CNN and transformer for medical image segmentation. *Phys. Med. Biol.* **68**, 175027 (2023).
33. Sang, S., Zhou, Y., Islam, M. T. & Xing, L. Small-object sensitive segmentation using across feature map attention. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 6289–6306 (2022).
34. Dong, R., Pan, X. & Li, F. DenseU-net-based semantic segmentation of small objects in urban remote sensing images. *IEEE Access* **7**, 65347–65356 (2019).
35. Song, I. & Kim, S. AVILNet: A new pliable network with a novel metric for small-object segmentation and detection in infrared images. *Remote Sens.* **13**, 555 (2021).
36. Huang, G., Ma, Y. & Yu, Y. Vehicle segmentation from remote sensing images using the small object segmentation convolutional network. In *2017 4th International Conference on Systems and Informatics (ICSAI)*. 1292–1296 (IEEE, 2017).
37. Jiang, Y. et al. APAUNet: Axis projection attention UNet for small target in 3D medical segmentation. In *Proceedings of the Asian Conference on Computer Vision*. 283–298 (2022).
38. He, W. et al. A statistical deformation model-based data augmentation method for volumetric medical image segmentation. *Med. Image Anal.* 102984 (2023).
39. Lou, A., Guan, S. & Loew, M. CaraNet: Context axial reverse attention network for segmentation of small medical objects. *J. Med. Imaging* **10**, 014005–014005 (2023).
40. Dijkstra, K., van de Loosdrecht, J., Atsma, W. A., Schomaker, L. R. & Wiering, M. A. CentroidNetV2: A hybrid deep neural network for small-object segmentation and counting. *Neurocomputing* **423**, 490–505 (2021).
41. Xie, L. et al. Recurrent saliency transformation network for tiny target segmentation in abdominal CT scans. *IEEE Trans. Med. Imaging* **39**, 514–525 (2019).
42. Liu, Z. et al. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10012–10022 (2021).
43. Sanchez-Garcia, R. et al. DeepEMhancer: A deep learning solution for cryo-EM volume post-processing. *Commun. Biol.* **4**, 874 (2021).
44. Badrinarayanan, V., Kendall, A. & Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481–2495 (2017).
45. Zhao, X. et al. M^2SNet: Multi-scale in Multi-scale Subtraction Network for Medical Image Segmentation. *arXiv preprint* arXiv:2303.10894 (2023).
46. Li, C. et al. Am-lfs: Automl for loss function search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8410–8419 (2019).
47. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1492–1500 (2017).
48. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **28**, 104863 (2020).
49. Codella, N. C. 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)*. 168–172 (IEEE, 2018).
50. Li, X., Wei, Y., Wang, L., Fu, S. & Wang, C. MSGSE-Net: Multi-scale guided squeeze-and-excitation network for subcortical brain structure segmentation. *Neurocomputing* **461**, 228–243 (2021).
51. Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M. & Asari, V. K. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. *arXiv preprint* arXiv:1802.06955 (2018).
52. Yang, M., Yu, K., Zhang, C., Li, Z. & Yang, K. DenseASPP for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
53. Qu, X. et al. EH-former: Regional easy-hard-aware transformer for breast lesion segmentation in ultrasound images. *Inf. Fusion* **109**, 102430. https://doi.org/10.1016/j.inffus.2024.102430 (2024).
54. Du, X. et al. UM-Net: Rethinking ICGNet for polyp segmentation with uncertainty modeling. *Med. Image Anal.* **99**, 103347. https://doi.org/10.1016/j.media.2024.103347 (2025).
55. Dai, D. et al. I2U-Net: A dual-path U-Net with rich information interaction for medical image segmentation. *Med. Image Anal.* **97**, 103241. https://doi.org/10.1016/j.media.2024.103241 (2024).
56. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4510–4520 (2018).
57. Chen, Y. et al. Dual path networks. In *Part of Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Vol. 30 (2017).
58. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778 (2016).
59. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*. 6105–6114 (PMLR, 2019).
60. Wang, W. et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8440–8449 (2019).
61. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).
62. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
63. Yang, Z., Zhu, L., Wu, Y. & Yang, Y. Gated channel transformation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
64. Lee, H., Kim, H.-E. & Nam, H. SRM: A style-based recalibration module for convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 62171133, in part by the Artificial Intelligence and Economy Integration Platform of Fujian Province, and the Fujian Health Commission, China under Grant 2022ZD01003.

Author contributions

Zhao.L Writing - Review & Editing, Conceptualization, Methodology, Software Wang.T responsible Visualization, Formal analysis. Chen.Y and Tang.H responsible Software, Visualization. Zhang.X and Tan.T responsible Data Curation, Resources Lin.F, Li.C and Li.Q responsible Data Curation, Validation Kang.D and Tong.T Writing - Review & Editing, Supervision, Funding acquisition.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.Z., D.K. or T.T.

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

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