Applied Intelligence (2025) 55:535
https://doi.org/10.1007/s10489-025-06448-8

QuinNet: Quintuple u-shape networks for scale- and shape-variant lesion segmentation

Gaojuan Fan[1] · Jie Wang[1] · Ruixue Xia[1] · Funa Zhou[2] · Chongsheng Zhang[1]

Accepted: 7 March 2025

Abstract

Deep learning approaches have demonstrated remarkable efficacy in medical image segmentation. However, they continue to struggle with challenges such as the loss of global context information, inadequate aggregation of multi-scale context, and insufficient attention to lesion regions characterized by diverse shapes and sizes. To address these challenges, we propose a new medical image segmentation network, which consists of one main U-shape network (MU) and four auxiliary U-shape sub-networks (AU), leading to Quintuple U-shape networks in total, thus abbreviated as *QuinNet* hereafter. MU devises special attention-based blocks to prioritize important regions in the feature map. It also contains a multi-scale interactive aggregation module to aggregate multi-scale contextual information. To maintain global contextual information, AU encoders extract multi-scale features from the input images, then fuse them into feature maps of the same level in MU, while the decoders of AU refine features for the segmentation task and co-supervise the learning process with MU. Overall, the dual supervision of MU and AU is very beneficial for improving the segmentation performance on lesion regions of diverse shapes and sizes. We validate our method on four benchmark datasets, showing that it achieves significantly better segmentation performance than the competitors. Source codes of QuinNet are available at https://github.com/Truman0o0/QuinNet.

Keywords Medical image segmentation · Multi-scale context aggregation · Global context information · Scale- and shape-variant

1 Introduction

Medical image analysis techniques play an increasing important role in clinical diagnosis, enabling medical staff to analyze various disease pathology more efficiently [1–4]. Medical image segmentation aims to automatically identify infection regions in medical images, which can substantially improve disease diagnosis efficiency [5]. In recent years, deep learning based methods have proven to be very effective in medical image segmentation and significant achievements have been made in this field.

The majority of deep learning-based medical image segmentation methods adopt an encoder-decoder deep neural network architecture, and the most representative and successful ones are UNet and its variants [6–9]. UNet [6] fuses low-level features in the encoder and high-level features in the decoder through skip connections. UNet++ [10] uses a set of dense skip connections to aggregate features of different scales. Attention UNet [11] utilizes an attention module within each skip connection to focus on the more important regions in a feature map.

Convolutional Neural Networks (CNNs) can learn local context information with low computational complexity, but have difficulty in capturing long-distance dependency. To overcome the limitations of CNNs, and inspired by the great success of Transformer [12] in the natural language processing (NLP) domain, researchers have tried to introduce Transformer into the vision domain, including medical image

✉ Ruixue Xia
xiaruixue@henu.edu.cn

✉ Chongsheng Zhang
cszhang@henu.edu.cn

Gaojuan Fan
fangaojuan@henu.edu.cn

Jie Wang
wangjie@henu.edu.cn

Funa Zhou
zhoufn@shmtu.edu.cn

[1] Henan University, Kaifeng 475001, China

[2] Shanghai Maritime University, Shanghai 200135, China

Published online: 15 March 2025

analysis. As an example, TransUNet [13] adopts a hybrid CNN-Transformer architecture in the encoder of the segmentation network, in which the feature map from CNN is tokenized into a 2D embedding and fed into a Transformer encoder to learn long-range context information.

In recent years, researchers have also attempted to combine CNN and all-MLP (MLP-based) [14, 15] architectures for medical image segmentation. UNeXt [16] adopts such a hybrid framework with an early convolutional stage and a MLP stage in the latent stage which tokenizes and projects the convolutional features and uses MLPs to model the representation. It can reduce computational complexity and the number of parameters while maintaining performance. In addition, there have also been attempts to use foundation models [17] or the Mamba architecture [18, 19] to more effectively or efficiently segment objects from medical images.

Despite the above advances, existing medical image segmentation methods still suffer from the following challenges. First, successive convolution operations make the model less effective in extracting global context information. Second, there lacks multi-scale context aggregation in the network. Third, the variety of shapes and sizes in the infection regions makes the current attention mechanism incapable in capturing different lesion areas precisely. Therefore, accurately segmenting lesion regions of diverse scales and shapes remains highly challenging.

To address the above challenges, we propose QuinNet, which is a new medical image segmentation network that consists of one main U-shape (MU) network and four or more auxiliary U-shape (AU) sub-networks, as depicted in Fig. 1. In the main U-shape (MU) network of QuinNet, we retain the U-shape encoder-decoder structure but devise two new modules, which are the feature selection extraction module (Attn-Block) and the multi-scale interactive aggregation module (MIA). The Attn-Blocks locate at the bottom of the encoder and the beginning of the decoder to enable the model to selectively focus on regions in the feature map that are more important for the segmentation task. The MIA module connects the encoder and the decoder and provides richer multi-scale context information. The auxiliary U-shape sub-networks of QuinNet (AU) help supervise the learning process. In AU, we first extract global context features of different scales from the input images, each level of feature map will then be fused with the corresponding feature map in the encoder of MU, employing a two-stage feature fusion block (TSF). The fused features at each level

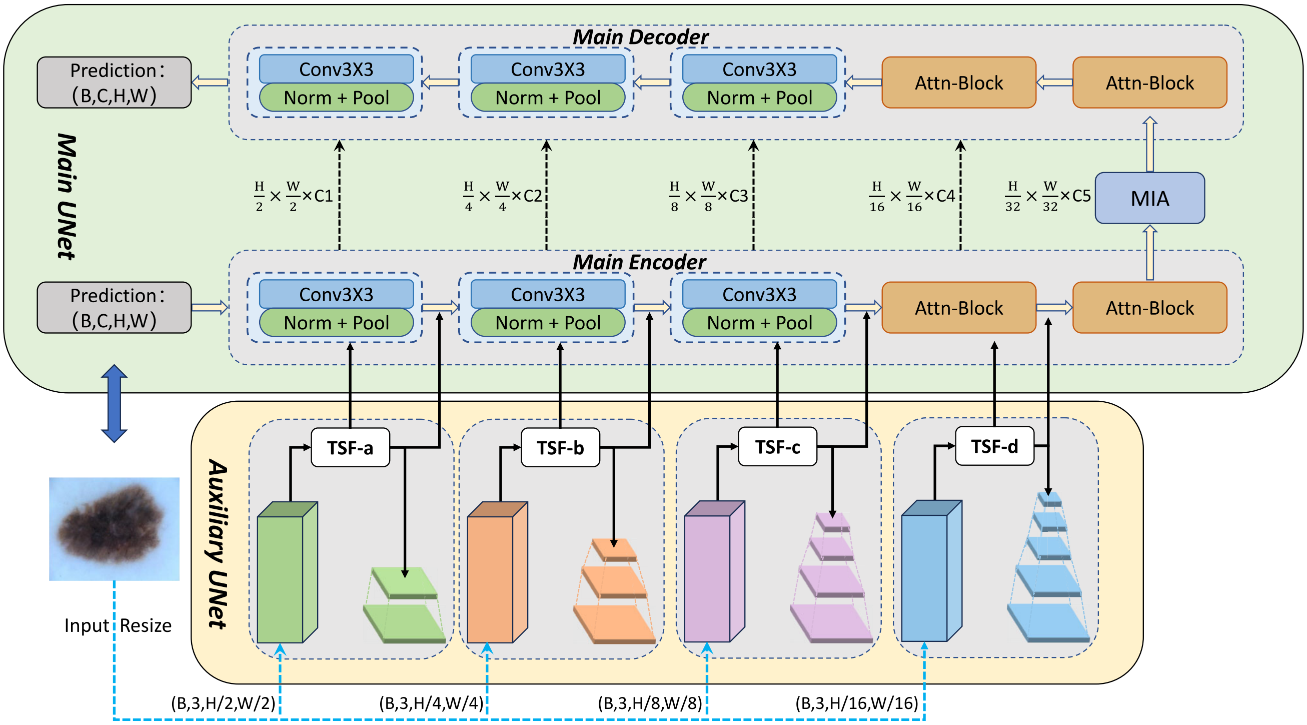


Fig. 1 Architecture of the proposed QuinNet method for scale- and shape-variant lesion region segmentation. QuinNet contains one main U-shape network and four auxiliary U-shape sub-networks. The top part depicts the main U-shaped (MU) network of QuinNet, which consists of a main encoder and decoder. We design the Attn-Block and MIA modules in MU to selectively focus on regions in the feature map that are more important for the segmentation task and aggregate multi-scale context information in the learning process. The bottom part depicts the auxiliary U-shape sub-networks of QuinNet, which aims at incorporating multi-scale global context information into the encoder of MU using the TSF block, and helping refine the attention in each level of features towards lesion region segmentation

will be simultaneously sent to the encoder of the MU and the decoder of the corresponding auxiliary sub-network to refine the features.

Overall, the encoders of AU alleviate the global context information loss problem, while the decoders of AU help refine different levels of features in MU and co-optimize the network with MU. The Attn-Blocks of MU distinguish the importance of different regions in the feature map, while the MIA module promotes multi-scale contextual information aggregation. Having AU and MU collaboratively supervise the learning process, the model can better preserve the global feature information and strengthen the multi-level semantic representation capability for scale- and shape-variant lesion segmentation.

To validate the performance of QuinNet, we conduct extensive experiments on four benchmark datasets, which are ISIC2018 [20], BUSI [21], CVC-ClinicDB [22], and the COVID-19 datasets. Experimental results demonstrate the performance improvements in lesion segmentation in comparison to state-of-the-art models. Moreover, in terms of computational efficiency/latency, it achieves comparable performance to lightweight segmentation models on the benchmark datasets.

The main contributions in this work can be summarized as follows:

1. We propose QuinNet, which is a new network architecture for scale- and shape-variant lesion region segmentation. It consists of one main U-shape (MU) network and four auxiliary U-shape (AU) sub-networks, which collaboratively supervise the learning process for lesion segmentation.
2. In the MU of QuinNet, we introduce Attn-Block and MIA modules to enable the model to focus on the important regions in the feature map and aggregate multi-scale context information, respectively.
3. Moreover, in the AU of QuinNet, we design four auxiliary U-shape sub-networks to strengthen the global context information in different levels of features in the encoder of MU and co-supervise the network with MU for the segmentation task.
4. We conduct extensive experiments on four medical image datasets, which demonstrate that QuinNet yields state-of-the-art performance.

The structure of this paper is as follows. Section 2 discusses related work in medical image segmentation using deep learning techniques. In Section 3, we present our proposed QuinNet architecture and the loss functions that we adopt. Section 4 reports the experimental results of our proposed method. Finally, Section 5 concludes the paper.

2 Related work

In this section, we will briefly review representative medical image segmentation methods that are categorized on the basis of the architectures they adopt, including Convolutional Neural Networks (CNNs), Vision Transformer, all-MLP (MLP-based), and Mamba architectures.

2.1 CNN-based segmentation techniques

The advent of Deep Neural Networks has revolutionized medical image processing, establishing them as indispensable tools for related tasks. UNet [6] was a seminal work in medical image segmentation, which adopts an encoder-decoder framework. UNet++ [10] utilizes hierarchical dense skip connections to bridge the semantic gap between features of the encoder and decoder. Attention UNet, a novel attention gate model, is introduced in [11], which selectively enhances target images with different shapes and sizes, while suppressing the area outside the lesion. UNet3+ [23] combines shallow details with deep semantics across different scales by using full-size skip connections and depth supervision.

PVT-MA [24] is a multi-attention-based model that contains a few attention modules to identify and locate polyps, and maximize the capture of polyp edge information, respectively. The authors in [25] present a superpixel-based masked image modeling method for skin lesion segmentation, in which they adopt Bayesian optimization for optimizing the superpixel generation and masking. AHGNN [26] presents a novel hypergraph-based network for medical image segmentation. It utilizes the degree of nodes to represent the connectivity among the nodes and constructs the hypergraph according to the size of objects, which can effectively capture the feature and local level information and is beneficial for enhancing segmentation performance. Liu et al. in [27] propose to incorporate the self-attention mechanism into UNet to enhance both the computational efficiency and receptive field interactions. They also design a content-aware decoder that can reassemble features based on predicted kernels to precisely rebuild image resolution.

[28] proposed the DermoSegDiff diffusion network [29] for skin lesion segmentation, which integrates texture, color, and boundary information into the segmentation task. The model assigns priority to the boundaries by gradually diminishing the importance of other regions during the training process, which ensures both lesion regions and healthy tissue are accurately delineated in the segmentation output.

As a new lightweight medical image segmentation model, TinyU-Net [30] devises a CMRF module which improves feature representation by fusing information from multi-receptive fields in a layer through cost-friendly cascaded depthwise separable convolutions, which is then used as a basic convolution block in UNet framework.

For few-shot medical image segmentaion, PMCR [31] devises a prototype correlation matching module to address false pairwise pixel correlation matches brought by large intra-class variations within the same class, and a class-relation reasoning module to explore inter-class relations to generalize to novel classes.

2.2 Transformer-based segmentation techniques

Motivated by Transformer's groundbreaking achievements in the NLP and vision tasks, researchers have endeavored to apply Transformer to medical image segmentation. Chen et al. [13] introduced TransUNet, which uses a CNN-Transformer hybrid architecture as the encoder of UNet, offering a robust alternative for medical image segmentation. Cao et al. [32] introduced SwinUNet, a pure Transformer-based method for medical image segmentation. Swin UNETR [33] is also a U-shaped network that contains a Transformer-based encoder and a CNN-based decoder for medical image analysis. Moreover, it utilizes various proxy tasks for self-supervised pretraining its encoder. The difference between TransUNet and Swin UNETR lies in that the former first employs a hybrid CNN-Transformer architecture as its encoder while the latter adopts a pure Transformer architecture in its encoder.

Yuan et al. in [34] propose a hybrid CNN and Transformer based network that produces complementary features which are then concatenated to fully utilize the advantages of both CNNs and Swin Transformers for more effective medical image segmentation. Similarly, in [35], the authors propose multi-modal model for lung lesions segmentation that combines a U-shape CNN branch for the extracting the visual features with another U-shaped vision Transformer branch for fusing the corresponding textual description features with the above visual features. Similarly, Tragakis et al. [36] proposed to leverage the capability of CNNs in learning powerful image representations and Transformers in capturing long-term multi-scale dependencies and obtained outstanding segmentation performance.

DB-SAM [17] is a dual-branch framework that adapts the Segment Anything Model (SAM) for high-quality medical image segmentaion. It contains a ViT branch and a convolution branch, in which the former first uses foundation model SAM to extract the initial feature representations, next incorporates a learnable channel attention block to capture domain-specific high-level features, while the latter employs a light-weight convolutional block to extract shallow features from the input medical image. Moreover, a bilateral cross-attention block is designed for effective cross-layer feature fusion between the two branches.

2.3 MLP-based segmentation techniques

MLPs used to have inferior performances to CNNs in computer vision tasks. To re-explore the potential of the MLPs in vision tasks, the authors in [14] propose MLP-Mixer, an all-MLP (MLP-based) architecture, that takes each image patch's embedding as input token and mixes channel and token information alternately.

UNeXt [16] is a hybrid framework for medical image segmentation which consists of an early convolutional stage and an MLP stage in the latent stage which tokenizes and projects the convolutional features and uses MLPs to model the representation. It can reduce computational complexity and the number of parameters while maintaining same or similar segmentation performance.

DPMNet [37] is an MLP-based network that includes a global and a local branch to understand the input images at different scales. In both branches, an axial residual connection MLP module is utilized to capture the long-range dependencies and local visual structures in the input image.

2.4 Mamba-based segmentation techniques

Recently, Mamba [18], a novel and efficient architecture, has emerged as a potential solution to address the constraints inherent in Transformer models, i.e., computational resource- and time-demanding. In comparison to Transformer architecture, Mamba achieves linear or near-linear scalability with sequence length while being able to capture long-range dependencies, and with significantly lower computational costs. Swin-UMamba [19] is the first attempt to discover the impact of pretrained Mamba-based networks in medical image segmentation, which adopts a U-shaped encoder-decoder framework with the encoder leveraging the power of pretrained Mamba-based foundation model.

In summary, UNet family and Transformers-based models are the mainstream methods for medical image segmentation and have demonstrated outstanding performance in recent years. However, their results come with a significant trade-off in terms of parameters and computational requirements, limiting their feasibility in resource-constrained environments. Hence, there is also growing interest in lightweight medical image segmentation models, including using Depthwise Separable Convolutions in CNN-based architectures, or employing the new MLP-based or Mamba-based architectures, as they are better suited for deployment in settings with limited resources.

3 Methodology

In this section, we first provide an overview of the QuinNet architecture, then describe the key modules in QuinNet in detail.

3.1 Overall architecture

The proposed QuinNet architecture is illustrated in Fig. 1, which consists of one main U-shaped network (MU for short) and four auxiliary U-shape sub-networks (AU for short). In a nutshell, MU includes a main encoder and decoder, while AU contains four pairs of encoder and decoder.

In the main encoder of MU, similar to UNet [6], each image will first undergo three successive convolution blocks. Next, specific Attn-Blocks are designed and put at the end of the main encoder and the beginning of the decoder of MU. With such Attn-Blocks, we can focus on regions in the feature map that are more important for the segmentation task. In addition, a multi-scale interactive aggregation (MIA) block is introduced between the encoder and decoder of MU to progressively fuse features of different receptive fields, aiming to aggregate multi-scale context information without losing resolution.

In the auxiliary U-shape sub-networks (AU), each encoder in the AU provides global context information of the input image at different scales, which will then be fused with the corresponding feature map of the same scale in the different encoder stages of MU, using the two-stage feature fusion block (TSF) that we devise. TSF blocks associate the global context feature from the AU-encoders with the successive encoding stages in MU to yield better feature representations. Next, each separate decoder in AU will evaluate the representation capability of different encoding stages in MU to help supervise the overall learning process. In brief, AU not only enhances feature representation capability of MU, but also captures the most useful features of different scales for the segmentation task and provides auxiliary supervision in the overall learning process of the model.

3.2 The MU module

The overall process of MU follows UNet, but we add two specific improvements, which are the Attn-Blocks and the MIA block, as can be seen from Fig. 2.

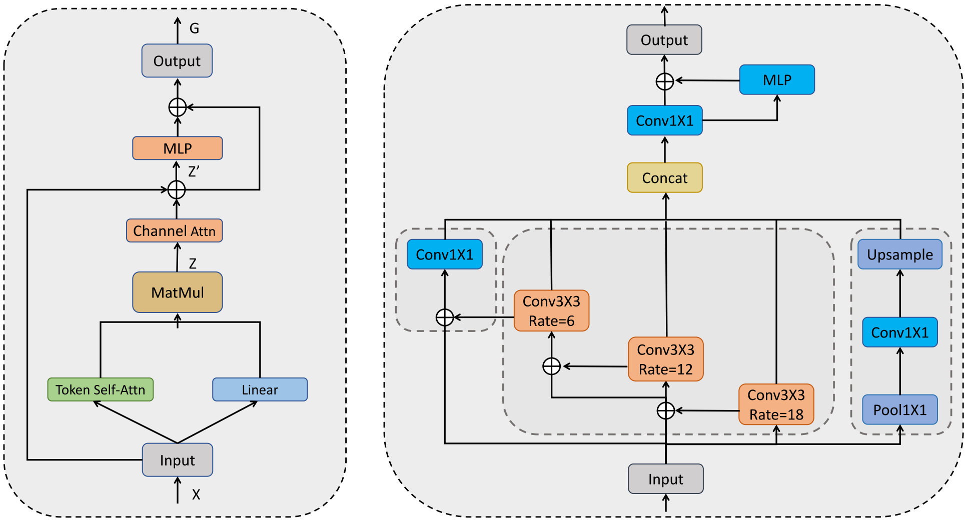
3.2.1 The attn-blocks

The Attn-Blocks are deployed after the successive convolution operations in the main encoder of MU, as well as in the beginning of the main decoder of MU. In the encoder of MU, the input of each Attn-Block is the fused feature map from a feature map from the preceding block in MU and the global feature of the input image at the same scale from AU, combined using the TSF block, which be detailed later. In the decoder of MU, the input of the first Attn-Block is the feature map obtained via the MIA block.

As shown in Fig. 2(a), Attn-Block is mainly based on self-attention and channel attention mechanisms. The input of the Attn-Block is subject to both Self-Attention (Token Self-Attn) and linear projections to extract more informative features using different mechanisms. A matrix multiplication is then performed between the two feature maps, and the new feature map will then undergo channel attention (Channel Attn) to focus on the more significant channels and further enhance the feature representation capability. Finally, we apply an MLP operation and residual connection to yield the resulting feature map. The detailed operations in Attn-Blocks can be formulated as follows:

$$Z = MatMul(SelfAttention(X), linear(X)) \tag{1}$$

$$Z^{'} = ChannelAttention(Z) + X \tag{2}$$



(a) The Attn-Blocks

(b) The MIA Block

Fig. 2 The Attn-Blocks and the MIA Block

$$G = MLP(Z^{'}) + Z^{'} \tag{3}$$

Overall, the Attn-Blocks leverage diverse attention mechanisms to prioritize and emphasize the more important channels and regions within the feature map, which can overcome the inadequate attention problem in the lesion regions and enable the network to learn stronger and more representative features.

3.2.2 The MIA block

As shown in Fig. 2(b), the MIA module connects the main encoder and decoder. Inspired by ASPP [38], MIA consists of three parts, including two point-wise convolutions (on the left and right, respectively) , as well as a dilated convolutions block. The two point-wise convolutions aim to increase the model's capacity by learning complex channel-wise interactions, while the dilated convolution employs receptive fields of varying scales to extract and aggregation multiple features, resulting in more robust semantic representations. Finally, the three features are aggregated.

Compared to ASPP, MIA takes a different approach by gradually fusing feature maps from neighboring scales using a residual connection, without simply concatenating feature maps from all scales, which can reduce the semantic gap between feature maps of different scales, leading to more representative global contextual feature representations, while still maintaining a relatively low computational cost as ASPP.

3.3 The AU module

Shown in Fig. 1, the AU module consists of quadruplet U-shape sub-networks, and their interactions and connections with the MU module is achieved through a specialized two-stage feature fusion block (TSF). In the following, we will elaborate on the AU encoder and decoder, and the TSF block.

3.3.1 The AU encoder and decoder

As can be observed from Fig. 1, in the encoder(s) of AU, we first resize the input image to four different scales (1/2, 1/4, 1/8, and 1/16), then extract the multi-scale features independently, using slightly different feature extraction blocks. For the first two scales (sub-encoders), we use 2 and 4 convolutional blocks (Conv-BN-ReLU), their corresponding outputs are then fed into TSF blocks, respectively. For the third scale (sub-encoder), besides the 4 convolutional blocks, we also add a self-attention block to capture the more important regions in the featuremap. Based upon the third sub-encoder, the fourth sub-encoder further incorporates a multi-scale convolution and feature fusion. Overall, we design the four sub-encoders in AU to capture the global feature/contextual information of different scales, which will then be separately integrated into feature maps of the same size in the encoder of MU to maintain the global context information in the successive convolutions in MU. Shown in Fig. 1, the AU-decoder is composed of four decoder sub-networks to supervise the feature learning of each scale, respectively. Each decoder sub-network contains two Conv-BN-ReLU blocks, and the segmentation loss is separately computed for each of them, which will be used to supervise the feature learning in AU and collectively guide the feature learning in the main encoder of MU.

In general, AU encoder and decoder provide auxiliary supervision in the overall learning process of MU, resulting in more contextual and representative features.

3.3.2 The TSF block

TSF connects AU with MU, shown in Fig. 3, it contains two fusion stages, the first stage mainly stacks and integrates the feature maps of two input sources, while the second stage aggregates the multi-scale contextual information using convolutions of different kernel sizes.

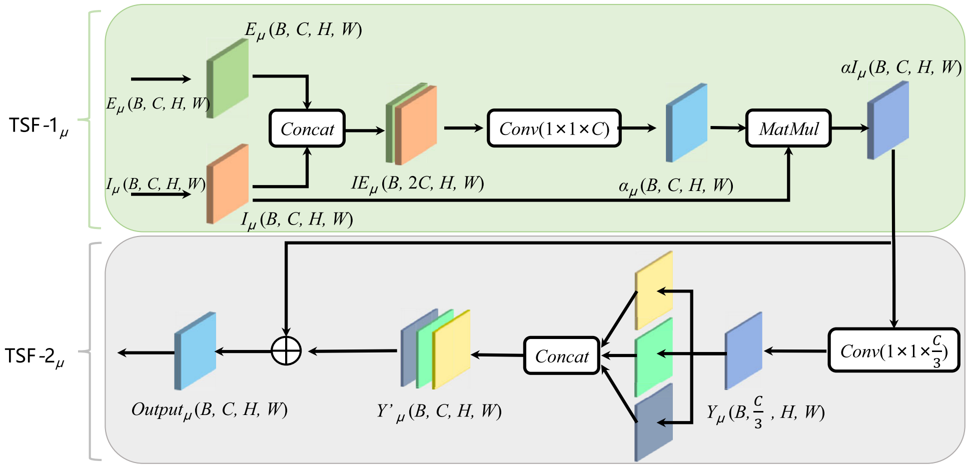


Fig. 3 The Two-stage Feature Fusion Block (TSF)

Specifically, TSF is divided into two stages, namely TSF-1_{μ} and TSF-2_{μ}, where $\mu \in (a, b, c, d)$. TSF-1_{μ} combines the scale-varying global feature information E_{μ} generated by each AU sub-encoder with the feature map of the same scale I_{μ} extracted by the encoder of MU. TSF-2_{μ} next employs the multi-scale feature fusion mechanism to further extract the contextual information in the fused feature map obtained via TSF-1_{μ}. The two stages can be formulated as follows.

$$IE_{\mu} = Concat(Conv_{1\times1\times C}(E_{\mu}), Conv_{1\times1\times C}(I_{\mu})) \tag{4}$$

$$\alpha I_{\mu} = MatMul(Relu(Conv_{1\times1\times C}(IE_{\mu})), Conv_{1\times1\times C}(I_{\mu})) \tag{5}$$

$$Y_{\mu} = Conv_{1\times1\times\frac{C}{3}}(\alpha I_{\mu}) \tag{6}$$

$$Y_{\mu}^{'} = Concat(Conv_{2\times2\times C}(Y_{\mu}), Conv_{4\times4\times C}(Y_{\mu}), Conv_{8\times8\times C}(Y_{\mu})) \tag{7}$$

$$Output_{\mu} = add(Y_{\mu}^{'}, \alpha I_{\mu}) \tag{8}$$

In summary, we devise the TSF block to more effectively incorporate the global contextual information learned via AU encoder(s) into the corresponding feature maps in the encoder of MU.

3.4 Loss Function

We utilize a combination of Binary Cross Entropy Loss (BCE Loss) and Dice Loss during the training process of QuinNet. The total loss $Loss_{total}$ comes from both MU and AU, the first two loss items calculate the segmentation loss in MU, while the last loss item denotes the segmentation loss in the quadruplet sub-networks of AU.

$$Loss_{total} = 0.5BCELoss(\hat{y}, y) + DiceLoss(\hat{y}, y) + \sum_{j=1}^{4} BCELoss(\hat{y_j}, y) \tag{9}$$

$$BCELoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(p_i)+(1-y_i)log(1-p_i)] \tag{10}$$

$$DiceLoss = 1 - \frac{2\left|y \cap \hat{y}\right|}{|y| + \left|\hat{y}\right|} \tag{11}$$

where N denotes batch size (number of samples), p_i denotes probability, $\hat{y}$ and y denote the prediction and ground-truth label, respectively.

4 Experiments

The experiments were conducted on a workstation that was equipped with an NVIDIA GeForce RTX 3070 GPU, operating on the Linux Mint OS. Our method is implemented using PyTorch 1.9.0 and Python 3.7.0. During network training, we employ the Adam optimizer with an initial learning rate of lr=0.0001, a batch size of 8, momentum set to 0.9, and 200 epochs.

4.1 Datasets and evaluation metrics

Datasets For the evaluation of the proposed approach, we use four publicly available medical image datasets, as described in Table 1. (1) the first dataset is ISIC2018 for skin lesion segmentation, which is a large-scale dataset of dermoscopy images published by the ISIC. It consists of 2594 training images with a size of 512×512. (2) the second BUSI dataset of breast cancer, which consist of 510 training images that are resized to 256×256. (3) CVC-ClinicDB is a commonly used dataset for polyp segmentation, including 550 high-resolution training images from 31 colonoscopy, each image has a size of 256×256. (4) the COVID-19 dataset consists of 3616 images with a size of 256×256.

We note that, for ISIC2018 and BUSI, we use the official splits for training and testing, and we collect the reported results from the corresponding papers. While for CVC-ClinicDB and COVID-19, since there are no official splits, we split the datasets on our own (with the ratio between the number of training and testing images being 9:1), then rerun all the competitors on the two datasets.

Table 1 Descriptions of the four benchmark datasets used in our experiments

Dataset	Images	Input size	Num of training	Num of tests
ISIC2018	2,594	512×512	2,594	100
BUSI	647	256×256	510	137
CVC-ClinicDB	612	256×256	550	62
COVID-19	3,616	256×256	3,255	361

Evaluation Metrics To evaluate the performance of different approaches, we used four evaluation metrics, which are Dice score (Dice), intersection over union (IoU), mean Intersection over Union (mIoU) and Recall. These evaluation metrics are formulated as follows:

$$Dice = 100 \times \frac{2TP}{2TP + FP + FN} \tag{12}$$

$$IoU = 100 \times \frac{TP}{TP + FP + FN} \tag{13}$$

$$mIoU = 100 \times \frac{1}{2} \times \left(\frac{TP}{TP + FP + FN} + \frac{TN}{TN + FN + FP}\right) \tag{14}$$

$$Recall = 100 \times \frac{TP}{TP + FN} \tag{15}$$

where TP, TN, FP and FN represent the number of True Positives, True Negatives, False Positives and False Negatives in the binary segmentation tasks, respectively.

Table 2 Comparison with other state-of-the-art methods on the four datasets

	Methods	Dice(%)	IoU(%)	mIoU(%)	Recall(%)
ISIC2018 dataset	UNet [6]	84.03	74.55	82.35	94.41
	Attention UNet [11]	88.99	80.32	86.34	92.97
	TransUNet [13]	88.78	82.63	-	85.78
	UNeXt [16]	89.70	81.70	85.38	92.39
	MTUNet [39]	89.84	81.66	87.28	96.18
	DermoSegDiff [28]	90.05	-	-	87.61
	SwinUnet [32]	88.69	79.96	85.88	96.96
	Swin UNETR [33]	90.40	82.68	87.79	96.72
	QuinNet(Ours)	**91.45**	**84.32**	**89.06**	**98.09**
BUSI dataset	UNet [6]	76.35	63.85	80.10	97.74
	UNet++ [10]	81.07	69.07	82.79	90.94
	Attention UNet [11]	81.75	69.93	83.43	89.44
	UNeXt [16]	79.37	66.95	81.70	89.20
	MTUNet [39]	80.66	68.27	82.47	88.95
	DermoSegDiff [28]	82.85	71.37	84.13	98.26
	SwinUnet [32]	81.98	69.88	83.34	98.19
	Swin UNETR [33]	84.71	73.96	85.91	98.33
	QuinNet(Ours)	**86.81**	**77.72**	**87.48**	**98.42**
CVC-ClinicDB dataset	UNet [6]	92.79	86.70	92.51	98.87
	UNet++ [10]	93.99	88.75	93.67	96.28
	Attention UNet [11]	94.06	88.92	93.76	96.29
	UNeXt [16]	90.20	82.42	90.07	93.90
	MTUNet [39]	92.42	86.50	92.40	95.00
	DermoSegDiff [28]	93.35	87.60	93.02	99.16
	SwinUnet [32]	93.59	86.74	92.53	97.53
	Swin UNETR [33]	94.21	88.61	93.46	98.44
	QuinNet(Ours)	**94.59**	**89.77**	**94.39**	**99.57**
COVID-19 dataset	UNet [6]	97.87	95.88	97.25	99.19
	UNet++ [10]	98.04	96.20	97.46	98.62
	Attention UNet [11]	97.73	95.63	97.45	98.63
	UNeXt [16]	97.45	95.08	96.72	98.20
	MTUNet [39]	97.85	95.85	97.22	98.51
	DermoSegDiff [28]	97.94	96.00	97.33	98.57
	SwinUnet [32]	97.89	95.92	97.25	99.01
	Swin UNETR [33]	98.07	96.23	97.52	99.26
	QuinNet(Ours)	**98.37**	**96.80**	**97.89**	**99.45**

The bold entries denote the best performance achieved on the corresponding dataset in terms of a specific evaluation metric

4.2 Results and analysis

Results on the ISIC2018 Dataset On ISIC2018, we compare our proposed QuinNet approach with UNet, Attention UNet, TransUNet, UNeXt, MTUNet, DermoSegDiff, SwinUnet, and Swin UNETR. As shown in Table 2, QuinNet outperforms the competitors by a large margin. In terms of the IoU metric, it outperforms the top-performing competitor by 1.69%. Moreover, the Dice, mIoU and Recall scores of QuinNet reach 91.45%, 89.06% and 98.09%, respectively, all of which significantly surpass the competitors. We visually compare the segmentation results of different models in Fig. 4(a).

Results on the BUSI Dataset On BUSI, we compare QuinNet with UNet, UNet++, Attention UNet, UNeXt, MTUNet, DermoSegDiff, SwinUnet, and Swin UNETR. The results are presented in Table 2. It is clear that our QuinNet method achieves the best performance, and the corresponding scores are 86.81%, 77.72%, 87.48% and 98.42%, in terms of Dice, IoU, mIoU and Recall, respectively. Compared to the best competitor, i.e., Swin UNETR, QuinNet achieves

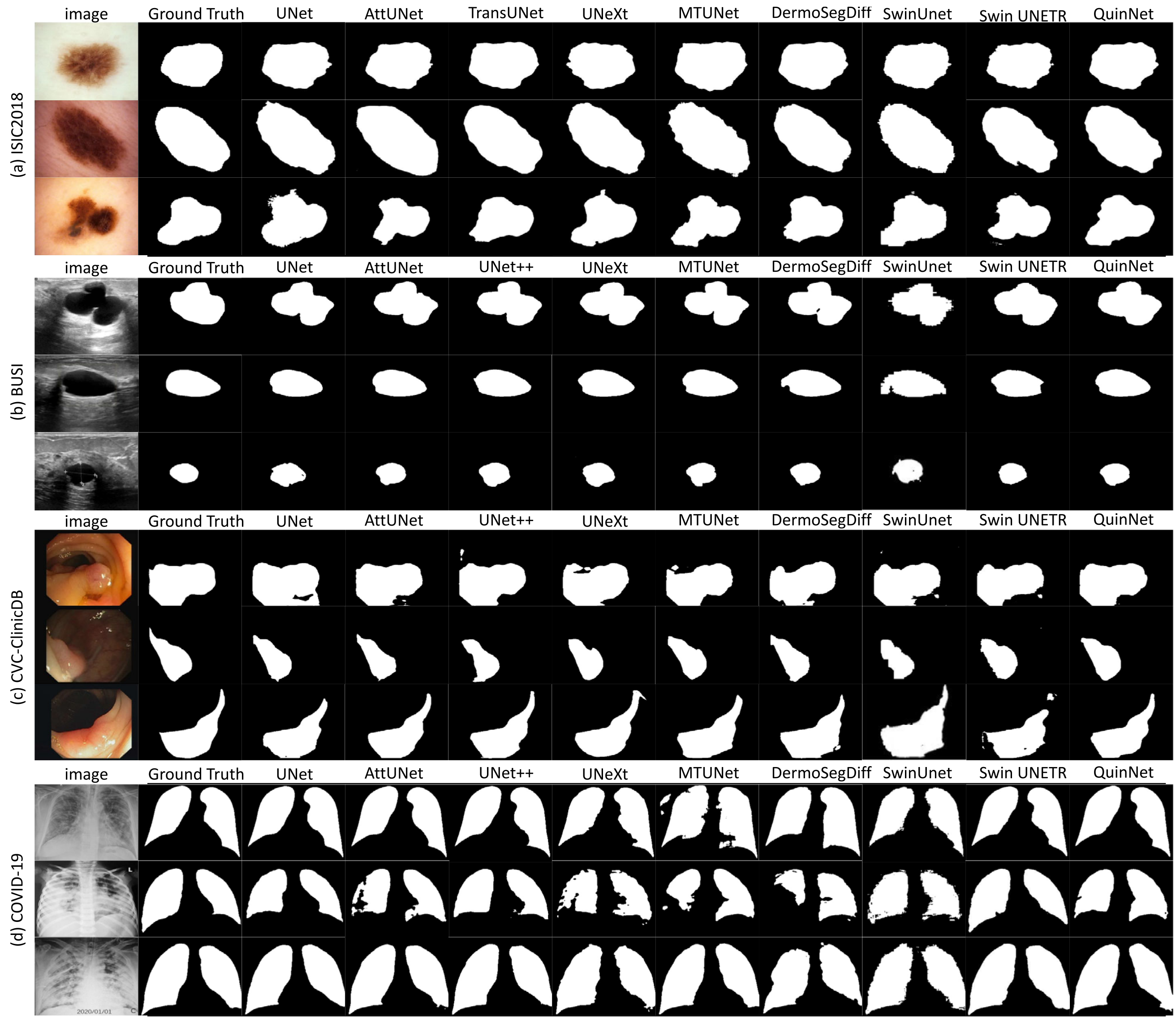


Fig. 4 Visual comparisons of different methods on the four datasets

Table 3 Comparison of parameter count(M), FLOPs(G), training Time(s) and inference time(ms) of different methods

Methods	Param (in M)↓	FLOPs (in G)↓	ISIC2018 Training Time (in s)	Inference Time (in ms)	BUSI Training Time (in s)	Inference Time (in ms)	CVC-ClinicDB Training Time (in s)	Inference Time (in ms)	COVID-19 Training Time (in s)	Inference Time (in ms)
UNet [6]	8.29	11.28	572.52	76.76	30.85	66.87	17.84	41.79	45.42	31.77
UNet++ [10]	9.63	27.55	680.55	49.87	82.80	86.35	19.60	39.69	115.9	39.58
Attention UNet [11]	8.42	11.60	571.22	55.46	31.99	76.19	19.28	36.14	53.36	35.98
TransUNet [13]	96.07	88.91	741.72	154.39	97.93	157.44	21.01	69.00	123.53	51.67
UNeXt [16]	1.47	0.57	547.91	25.10	11.91	34.39	6.14	24.86	34.32	11.71
MTUNet [39]	79.07	42.22	673.45	69.41	86.64	77.65	17.47	48.07	80.83	53.49
DermoSegDiff [28]	61.10	38.13	685.39	73.47	83.57	77.15	16.67	43.89	65.72	46.86
SwinUnet [32]	27.15	5.91	565.77	51.96	24.74	29.85	18.73	29.89	49.58	39.58
Swin UNETR [33]	6.29	3.65	595.95	107.56	20.22	91.79	20.26	43.45	58.53	43.91
QuinNet(Ours)	7.49	2.12	565.87	50.31	19.60	43.75	12.15	37.82	56.30	38.03

improvements of 2.1% and 1.16% in terms of Dice and IoU, respectively, which are substantial. Some of the segmentation results are demonstrated in Fig. 4(b).

Results on the CVC-ClinicDB Dataset CVC-ClinicDB is a commonly used segmentation dataset. Table 2 demonstrates the performance of QuinNet in polyp lesion segmentation, which yields the best performance, with an accuracy of 94.59% (in terms of Dice), 89.77%(in terms of IoU), 94.39%(in terms of mIoU) and 99.57%(in terms of Recall). In Fig. 4(c), we showcase results of QuinNet alongside other approaches, demonstrating that QuinNet delivers significantly improved segmentation performance in comparison to the competitors.

Results on the COVID-19 Dataset Table 2 also shows the results of the proposed method and the competing methods on COVID-19. We can see that all the methods tend to saturate their performance in segmentation, since the segmentation areas in COVID-19 are large relatively, and the shapes of the segmentation areas are generally invariant. As a result, COVID-19 does not present scale- and shape-variant lesion segmentation challenges as other datasets. We compare our method with the existing methods on this dataset

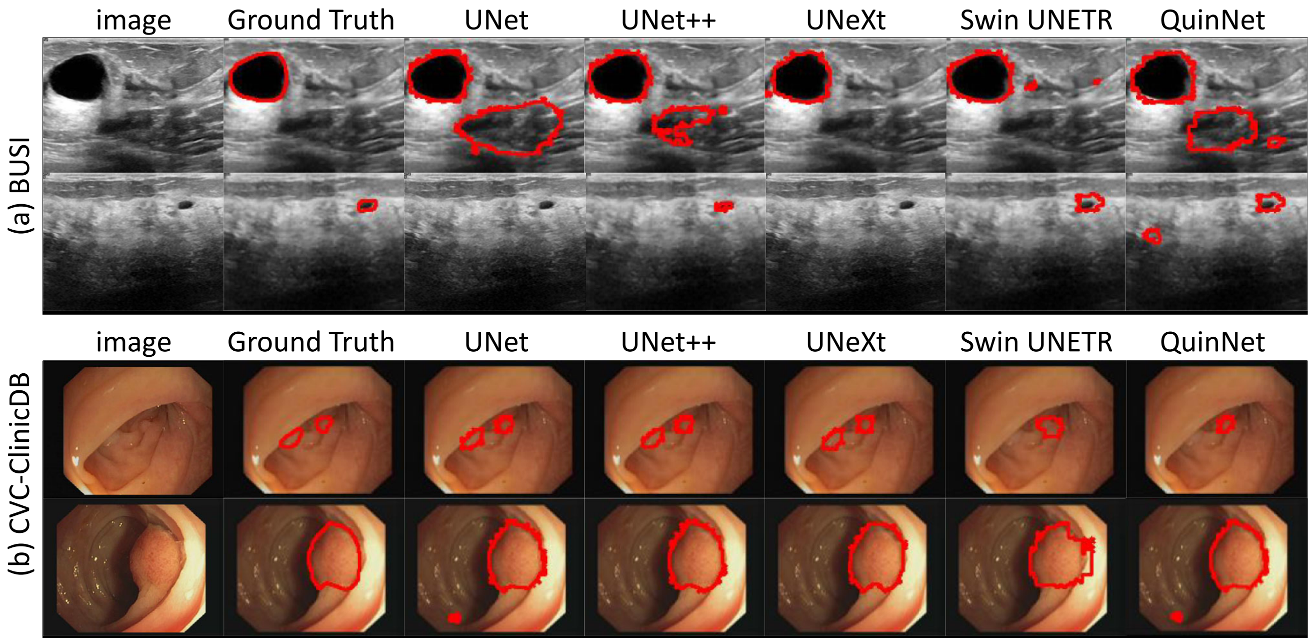


Fig. 5 Visualization of failure cases on the BUSI and CBC-ClinicDB datasets

Table 4 Ablation study on the effects of different modules in QuinNet

Modules					ISIC2018		BUSI		CVC-ClinicDB	
Baseline	Attn-Block	MIA	AU	TSF	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)
✓					89.70	81.70	79.37	66.95	90.20	82.42
✓	✓				88.87	80.15	81.50	69.47	91.03	83.94
✓	✓	✓			90.17	82.19	84.24	73.14	92.07	85.35
✓	✓	✓	✓		90.26	82.47	83.22	71.98	93.22	87.45
✓	✓	✓	✓	✓	**91.45**	**84.32**	**86.81**	**77.72**	**94.59**	**89.77**

The bold entries denote the best performance achieved on the corresponding dataset in terms of a specific evaluation metric

to demonstrate that QuinNet can still yield excellent performance on benchmark datasets that do not present scale- and shape-variant lesion characteristics. Overall, on COVID-19, our method obtains a Dice score of 98.37%, an IoU score of 96.80%, an mIoU score of 97.89% and a Recall of 99.45%, which surpass all the competitors once again. Some visualization results are shown in Fig. 4(d).

Efficiency Performance Since our model has one main UNet and five auxiliary UNets, it is necessary to check the efficiency performance of our model, in order to eliminate the concern that it may be time-consuming in training. In Table 3, we report the parameter count(M), FLOPs(G), training time(s) and inference time(ms) of different methods. We see that, UNeXt is the most efficient algorithm, followed by QuinNet (ours), SwinUnet and Swin UNETR. In specific, on ISIC2018, the training and inference efficiency of QuinNet is almost the same as SwinUnet, while on BUSI and CVC-ClinicDB, QuinNet is more efficient in training than SwinUnet, but the latter outperforms the former in inference efficiency. Overall, the efficiency of our proposed QuinNet model is on par with SwinUnet. However, Since QuinNet outperforms both SwinUnet and Swin UNETR in segmentation performance, as analyzed above, it is more preferable than SwinUnet and Swin UNETR in practice, when considering both the segmentation performance and model efficiency simultaneously.

Visual analysis As shown in Fig. 4, we provide visual comparisons of various methods.

(1) Visual evaluations of various models on the ISIC2018 dataset reveal that our method exhibits enhanced robustness in handling intricate segmentation shapes. This improvement can be attributed to the introduction of Attn-Blocks and the MIA block in our approach.
(2) Visual evaluations of various models on the BUSI dataset. The results clearly illustrate that our proposed approach excels in detecting scale-variant lesions compared to its counterparts, owing to the fact that MU integrates multi-scale global contextual information extracted by the sub-encoders of AU.

Table 5 Ablation study on the design of AU

	AU Structures	Dice(%)	IoU(%)	mIoU(%)	Recall(%)
	AU-None	90.17	82.19	87.59	93.24
ISIC2018	AU-RandomCrop	90.95	83.50	88.52	96.94
dataset	AU-CropTarget-Resize	76.11	61.93	71.78	94.04
	AU-Resize	90.93	83.44	88.50	97.20
	AU-Resize-Supervision(Ours)	**91.45**	**84.32**	**89.06**	**98.09**
	AU-None	84.24	73.14	84.68	90.08
BUSI	AU-RandomCrop	84.80	74.17	85.74	98.53
dataset	AU-CropTarget-Resize	57.58	41.19	65.08	**98.81**
	AU-Resize	82.98	71.64	84.28	98.27
	AU-Resize-Supervision(Ours)	**86.81**	**77.72**	**87.48**	98.42
	AU-None	92.07	85.35	91.71	95.67
CVC-ClinicDB	AU-RandomCrop	93.16	87.23	92.81	99.28
dataset	AU-CropTarget-Resize	66.71	50.45	69.93	99.18
	AU-Resize	92.55	86.29	92.28	99.03
	AU-Resize-Supervision(Ours)	**94.59**	**89.77**	**94.39**	**99.57**

The bold entries denote the best performance achieved on the corresponding dataset in terms of a specific evaluation metric

Table 6 The impact of different numbers of AU-encoders on QuinNet

AU sub-networks	ISIC2018 Dice(%)	IoU(%)	BUSI Dice(%)	IoU(%)
1	90.14	82.17	84.28	73.44
2	90.46	83.14	84.12	73.08
3	90.10	82.10	83.71	72.52
4	**91.45**	**84.32**	**86.81**	**77.72**

The bold entries denote the best performance achieved on the corresponding dataset in terms of a specific evaluation metric

(3) Visual evaluations of various models on the CVC-ClinicDB dataset. The results once again highlight the effectiveness of our approach in accurately segmenting lesions of diverse shapes and sizes.

(4) Visual evaluations of various models on the COVID-19 dataset. We see that our segmentation performance remains the best on COVID-19.

Failure cases We also provide a few failure cases where the best competitor outperforms our method. In Fig. 5, we observe that, comparing to the best competitor Swin UNETR, our proposed method sometimes misses the true positive results, which may be partially due to the confusing lesion areas in the original image. For instance, in the third image of Fig. 5, the left ground-truth poly area has inherent confusion. We also notice that our method may produce false positive lesion prediction results, as can be seen from the other three images in Fig. 5. Therefore, certain inherent feature filtering mechanism or post-processing or ensemble learning steps need to be incorporated into QuinNet to reduce such false positive predictions.

4.3 Ablation study

In order to evaluate the influence of different modules in QuinNet, ablation studies are performed on ISIC2018, BUSI, and CVC-ClinicDB.

Ablation study on the effects of different modules in QuinNet To validate the effects of different modules in QuinNet, we constantly combine different modules using UNet as the baseline. As shown in Table 4, we progressively add Attn-Block, MIA, AU and TSF to the baseline, denoted as Baseline+Attn-Block, Baseline+Attn-Block+MIA, Baseline+Attn-Block+MIA+AU and Baseline +Attn-Block+MIA+AU+TSF(Ours), respectively. It can be observed that QuinNet's performance has improved to a certain extent after the incorporation of various modules. These specially designed modules enable the preservation of global contextual information and the identification of critical channels and regions in the feature maps.

Ablation study on the design of AU Besides the current AU-encoder and AU-decoder structures, we also tested other possible network structures for AU. AU-None indicates removing the AU module, AU-RandomCrop and AU-Resize denote using random crop and straightforward resize operations for AU-encoder, respectively. AU-CropTarget-Resize means first cropping the target areas then resizing them, while AU-Resize-Supervision represents our method. The results of the comparison are shown in Table 5, AU in QuinNet is generally an improvement in model performance. In the meantime, for AU-CropTarget-Resize, we observed low segmentation accuracy, possibly due to the fact that it only focuses on the lesion regions but ignores the global context, leading to its incapability in shape- and scale-variant lesion segmentation.

Ablation study on the number of AU sub-networks To verify the effect of the number of sub-networks in the AU encoder on QuinNet, we are also conducting ablation studies. The results of the experimental comparison are shown in Table 6, we see that quadruplet AU-encoders generally yield the best performance on ISIC2018 and BUSI, since it can learn more abundant multi-scale global contextual information.

Ablation study on the weight parameter for BCE loss In Table 7, we check the influence of the weight parameter for BCE loss on model performance, in which we test 5 different values including 0.125, 0.25, 0.5, 1 and 2. It is clear that, when the weight parameter sets to 0.5, our model obtains the

Table 7 The impact of the weight parameter for BCE loss on model performance

Weight for BCE loss	ISIC2018 Dice(%)	IoU(%)	BUSI Dice(%)	IoU(%)	CVC-ClinicDB Dice(%)	IoU(%)
0.125	90.23	82.56	83.74	73.44	93.91	88.55
0.25	89.83	81.91	84.61	73.73	94.32	89.17
1	90.48	83.06	84.41	73.93	94.50	89.49
2	89.73	81.62	83.49	73.20	92.95	86.89
0.5	**91.45**	**84.32**	**86.81**	**77.72**	**94.59**	**89.77**

The bold entries denote the best performance achieved on the corresponding dataset in terms of a specific evaluation metric

Table 8 The impact of Attn-Block on model performance

Num of Attn-Block		ISIC2018		BUSI		CVC-ClinicDB	
Encoder	Decoder	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)
3	3	91.26	84.05	**86.81**	**77.72**	94.58	89.75
1	1	89.65	81.59	85.75	76.08	92.19	85.61
1	2	91.29	84.08	86.57	77.54	92.60	86.29
2	0	90.61	83.00	85.34	75.50	92.73	86.49
2	1	90.76	83.22	86.44	77.16	92.71	86.48
2	2	**91.45**	**84.32**	84.73	73.95	**94.59**	**89.77**

The bold entries denote the best performance achieved on the corresponding dataset in terms of a specific evaluation metric

best performance. Therefore, we set the default value of the weight parameter for BCE loss to be 0.5 in other experiments.

Ablation study on the number of Attn-Blocks in the encoder and decoder of MU In the main U-shape network (MU), we devise the Attn-Blocks in the encoder and decoder. In this ablation study, we study the influence of the number of Attn-Blocks on the performance of QuinNet, where we vary different number of Attn-Blocks in the encoder and decoder. It can be observed from Table 8 that, on BUSI, the number of Attn-Blocks has a remarkable influence on the model performance, when we change the numbers of Attn-Blocks in the encoder and decoder from (2,2) to (3,3), there is a significant performance improvement by more than 2%. We also did experiments on the COVID-19 dataset (which were not included due to table size limit), when this hyperparameter changes from (2,2) to (1,1), model performance improves from 98.01 (Dice) and 96.14 (IoU) to 98.37 and 96.80, respectively. Therefore, we need to tune this hyperparameter separately on different datasets.

5 Conclusion

In this paper, we present the QuinNet approach to lesion segmentation. It consists of one main U-shape network (MU) and four auxiliary U-shape sub-networks (AU). The Attn-Blocks and the MIA block in MU help the model focus on the important regions in the feature map and capture multi-scale context information, while AU utilizes auxiliary U-shape sub-networks to capture and maintain global contextual information at multiple scales, it also provides auxiliary supervision in the overall learning process of the model. Extensive experiments show that QuinNet obtains very outstanding performance in shape- and scale-variant lesion segmentation. One limitation of QuinNet is that it may produce false positive predictions. In future work, we will incorporate feature filtering mechanism into QuinNet to reduce such false positive results. We will also adopt lightweight techniques such as depthwise separable convolutions to further improve the efficiency of our model.

Acknowledgements This paper was partially supported by the Henan Provincial Science and Technology Projects (No.232102211021 and No.232102311012), Henan Provincial Health Department Medical Science and Technology Research Plan (SBGJ202103098), Henan Provincial Center for Outstanding Overseas Scientists (No.GZS2025004), Henan Provincial High-Level Talent International Training Program (No.GCC2025010), and the National Natural Science Foundation of China (No. 82100052).

Author Contributions Gaojuan Fan: Supervision, Investigation, Methodology, Writing- Reviewing and Editing. Jie Wang: Conceptualization, Methodology, Software, Writing- Original draft preparation. Ruixue Xia: Investigation, Methodology, Writing- Reviewing and Editing. Funa Zhou: Writing- Reviewing and Editing. Chongsheng Zhang: Supervision, Investigation, Methodology, Writing- Reviewing and Editing.

Data Availability The data that support the findings of this study are publicly available in [ISIC2018 dataset, BUSI dataset, and CVC-ClinicDB dataset].

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical and informed consent for data used This paper was done by the authors, and no human participants other than the authors were involved in it, and informed consent was obtained from all authors.

References

1. Dayarathna S, Islam KT, Uribe S, Yang G, Hayat M, Chen Z (2024) Deep learning based synthesis of MRI, CT and PET: review and analysis. Med Image Anal 92:103046
2. Xiao Y, Chen C, Fu X, Wang L, Yu J, Zou Y (2023) A novel multi-task semi-supervised medical image segmentation method based on multi-branch cross pseudo supervision. Appl Intell 53(24):30343–30358

3. Liu H, Xu Z, Gao R, Li H, Wang J, Chabin G, Oguz I, Grbic S (2024) Cosst: Multi-organ segmentation with partially labeled datasets using comprehensive supervisions and self-training. IEEE Trans Med Imaging
4. Goel T, Murugan R, Mirjalili S, Chakrabartty DK (2021) Optconet: an optimized convolutional neural network for an automatic diagnosis of COVID-19. Appl Intell 51(3):1351–1366
5. Lê M, Unkelbach J, Ayache N, Delingette H (2015) GPSSI: gaussian process for sampling segmentations of images. In: International conference on medical image computing and computer-assisted intervention (MICCAI), vol 9351, pp 38–46
6. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention (MICCAI), vol 9351, pp 234–241
7. Kugelman J, Allman J, Read SA, Vincent SJ, Tong J, Kalloniatis M, Chen FK, Collins MJ, Alonso-Caneiro D (2022) A comparison of deep learning u-net architectures for posterior segment oct retinal layer segmentation. Sci Rep 12(1):14888
8. Zhu L, Zhan S, Zhang H (2019) Stacked u-shape networks with channel-wise attention for image super-resolution. Neurocomputing 345:58–66
9. Li A, Li X, Ma X (2024) Residual dual u-shape networks with improved skip connections for cloud detection. IEEE Geosci Remote Sens Lett 21:1–5
10. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018) Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop (DLMIA), vol 11045, pp 3–11
11. Oktay O, Schlemper J, Folgoc LL, Lee MCH, Heinrich MP, Misawa K, Mori K, McDonagh SG, Hammerla NY, Kainz B, Glocker B, Rueckert D (2018) Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems (NeurIPS), pp 5998–6008
13. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y (2021) Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306
14. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J, Lucic M, Dosovitskiy A (2021) Mlp-mixer: An all-mlp architecture for vision. In: Annual conference on neural information processing systems (NeurIPS), pp 24261–24272
15. Touvron H, Bojanowski P, Caron M, Cord M, El-Nouby A, Grave E, Izacard G, Joulin A, Synnaeve G, Verbeek J, Jégou H (2023) Resmlp: Feedforward networks for image classification with data-efficient training. IEEE Trans Pattern Anal Mach Intell 45(4):5314–5321
16. Valanarasu JMJ, Patel VM (2022) Unext: Mlp-based rapid medical image segmentation network. In: International conference on medical image computing and computer-assisted intervention (MICCAI), vol 13435, pp 23–33
17. Qin C, Cao J, Fu H, Khan FS, Anwer RM (2024) Db-sam: Delving into high quality universal medical image segmentation. In: International conference on medical image computing and computer-assisted intervention (MICCAI), pp 498–508
18. Gu A, Dao T (2023) Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752
19. al JL (2024) Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International conference on medical image computing and computer-assisted intervention (MICCAI), pp 615–625
20. Codella NCF, Gutman DA, Celebi ME, Helba B, Marchetti MA, Dusza SW, Kalloo A, Liopyris K, Mishra NK, Kittler H, Halpern A (2018) Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi). In: 15th IEEE international symposium on biomedical imaging (ISBI), pp 168–172
21. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A (2020) Dataset of breast ultrasound images. Data Brief 28:104863
22. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Miguel CR, Vilariño F (2015) WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput Med ImagingGraph 43:99–111
23. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen YW, U-net JW (2004) A full-scale connected unet for medical image segmentation. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1055–1059
24. Shang X, Wu S, Wang S (2025) Pvt-ma: pyramid vision transformers with multi-attention fusion mechanism for polyp segmentation. Appl Intell 55. https://doi.org/10.1007/s10489-024-06041-5
25. Wang Z, Lyu J, Tang X (2023) autosmim: Automatic superpixel-based masked image modeling for skin lesion segmentation. IEEE Trans Med Imaging 42(12):3501–3511
26. Chai S, Jain RK, Mo S, Liu J, Yang Y, Li Y, Tateyama T, Lin L, Chen YW (2024) A novel adaptive hypergraph neural network for enhancing medical image segmentation. In: International conference on medical image computing and computer-assisted intervention (MICCAI 2024), pp 23–33
27. Liu X, Gao P, Yu T, Wang F, Yuan R (2025) Cswin-unet: Transformer unet with cross-shaped windows for medical image segmentation. Inf Fusion 113
28. Bozorgpour A, Sadegheih Y, Kazerouni A, Azad R, Merhof D (2023) Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation. In: International workshop on predictive intelligence in medicine (PRIME), vol 14277, pp 146–158
29. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Annual conference on neural information processing systems (NeurIPS)
30. Chen J, Chen R, Wang W, Cheng J, Zhang L, Chen L (2024) Tinyu-net: Lighter yet better u-net with cascaded multi-receptive fields. In: International conference on medical image computing and computer-assisted intervention (MICCAI), pp 626–635
31. Zhang Y, Li H, Gao Y, Duan H, Huang Y, Zheng Y (2024) Prototype correlation matching and class-relation reasoning for few-shot medical image segmentation. IEEE Trans Med Imaging
32. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M (2022) Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV 2022 Workshops, pp 205–218
33. Tang Y, Yang D, Li W, Roth HR, Landman B, Xu D, Nath V, Hatamizadeh A (2022) Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2022), pp 20730–20740
34. Yuan F, Zhang Z, Fang Z (2023) An effective CNN and transformer complementary network for medical image segmentation. Pattern Recogn 136:109228
35. Li Z, Li Y, Li Q, Wang P, Guo D, Lu L, Jin D, Zhang Y, Hong Q (2024) Lvit: Language meets vision transformer in medical image segmentation. IEEE Trans Med Imaging 43(1):96–107
36. Tragakis A, Kaul C, Murray-Smith R, Husmeier D (2023) The fully convolutional transformer for medical image segmentation. In: IEEE/CVF winter conference on applications of computer vision (WACV), pp 3649–3658
37. Wang S, Zhao X, Zhang Y, Zhao Y, Zhao Z, Ding H, Chen T, Qiao S (2024) Dpmnet: Dual-path mlp-based network for aneurysm

image segmentation. In: International conference on medical image computing and computer-assisted intervention (MICCAI), pp 245–254

38. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
39. Wang H, Xie S, Lin L, Iwamoto Y, Han X, Chen Y, Tong R (2022) Mixed transformer u-net for medical image segmentation. In: IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 2390–2394

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