Signal, Image and Video Processing
https://doi.org/10.1007/s11760-023-02528-9

ORIGINAL PAPER

N-Net: an UNet architecture with dual encoder for medical image segmentation

Bingtao Liang[1] · Chen Tang[1] · Wei Zhang[2] · Min Xu[1] · Tianbo Wu[1]

Received: 10 March 2022 / Revised: 8 July 2022 / Accepted: 7 February 2023

Abstract
In order to assist physicians in diagnosis and treatment planning, accurate and automatic methods of organ segmentation are needed in clinical practice. UNet and its improved models, such as UNet + + and UNt3 + , have been powerful tools for medical image segmentation. In this paper, we focus on helping the encoder extract richer features and propose a N-Net for medical image segmentation. On the basis of UNet, we propose a dual encoder model to deepen the network depth and enhance the ability of feature extraction. In our implementation, the Squeeze-and-Excitation (SE) module is added to the dual encoder model to obtain channel-level global features. In addition, the introduction of full-scale skip connections promotes the integration of low-level details and high-level semantic information. The performance of our model is tested on the lung and liver datasets, and compared with UNet, UNet + + and UNet3 + in terms of quantitative evaluation with the Dice, Recall, Precision and F1 score and qualitative evaluation. Our experiments demonstrate that N-Net outperforms the work of UNet, UNet + + and UNet3 + in these three datasets. By visual comparison of the segmentation results, N-Net produces more coherent organ boundaries and finer details.

Keywords Deep learning · Image segmentation · Encoder–decoder · Convolutional neural network

1 Introduction

Medical research produces a large number of medical images [1], mainly including computed tomography (CT), magnetic resonance imaging (MRI), ultrasound imaging, and more. The purpose of medical image segmentation is to segment the parts with some special meanings, extract relevant features and provide reliable basis for clinical diagnosis and pathology research. With the successful application of medical images in clinical medicine, image segmentation is playing a more and more important role in medical images. Because medical images have a series of problems such as inhomogeneity and individual difference, manual annotation is still used in clinical practice. This work is time-consuming and requires experienced specialists to complete. Therefore, accurate and reliable automatic segmentation method is in high demand, which can reduce the workload of clinical medical experts and improve efficiency.

Recently, convolutional neural networks (CNNs) have achieved advanced performance in a wide range of visual recognition tasks [2–6]. Jonathan et al. began to try to use CNNS to complete the end-to-end automatic segmentation tasks and proposed fully convolutional neural network (FCN) [7]. FCN popularized the use of end-to-end CNNs in image segmentation. The main contributions of FCN are as follows: FCN used the convolutional layer instead of the full connection layer to obtain image spatial information as much as possible; the deconvolution layer was used to upsample the feature images to obtain the segmented image which meets the size requirement. Since FCN, CNNs have been widely used in the field of image segmentation, which greatly promotes to develop many segmentation models, e.g., PSPNet [8], RefineNet [9] and a series of DeepLab version [10–12], UNet [13] and so on. Recently, there were many new applications of FCN to medical image segmentation. Tong et al. [14] proposed a shape representation model constrained fully convolutional neural networks. This model combined a fully convolutional neural network with a shape representation

✉ Wei Zhang
zhangwei5660@126.com

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

[2] Tianjin Key Laboratory of Ophthalmology and Visual Science, Tianjin Eye Institute, Clinical College of Ophthalmology of Tianjin Medical University, Tianjin Eye Hospital, Tianjin 300020, China

Published online: 22 March 2023

model (SRM) to segment nine organs in the head and neck from computed tomography (CT) images. Wang et al. [15] proposed a dual-input V-mesh fully convolutional network to segment the pancreas in abdominal CT images. They simultaneously sent the original CT scans and images processed by a contrast-specific graph-based visual saliency (GBVS) algorithm to the network to improve the contrast of the pancreas and other soft tissues. Xue et al. [16] proposed a novel, automatic multiorgan segmentation algorithm based on a new hybrid neural-like P system to achieve the automatic segmentation of organs-at-risk. These works have yielded good results, but there is still room for improvement.

Medical images have two characteristics: particularity and rarity [17, 18]. Particularity refers to the blurring of the boundary of medical images. Rarity refers to the fact that pixel-level annotation data of medical images are extremely rare and difficult to obtain. UNet, which is based on an encoder–decoder architecture [19], has been widely used in medical image segmentation. UNet combined low-resolution information with high-resolution information to accurately locate and identify organs. And UNet has fewer parameters than FCN(VGG16 Backbone)[7], so high-quality models can be trained with fewer medical images. Many image segmentation studies [20–22] showed that feature maps of different scales contain unique image information. Further, these multi-context representations [11, 19, 20, 23] are manually designed, lacking flexibility to model the multi-context representations. This makes that long-range object relationships in the whole images cannot be fully leveraged in these approaches [11, 14, 20, 23], which is pivotal importance in many medical imaging segmentation problems. UNet solves the problem of information loss and feature information fusion by designing skip connections. The skip connections of UNet ensure that the finally recovered feature maps integrate more low-level features. Although UNet is suitable for medical image segmentation, its segmentation accuracy on many datasets still has much room for improvement. These improved UNet models dominate the literature in medical image segmentation and have achieved outstanding performance in a broad span of applications, such as brain [24] or cardiac imaging [25].

In order to recede the fusion of semantically different features from the ordinary skip connections in UNet, UNet + + [26] further strengthens ordinary skip connections by introducing nested and dense skip connections. Moreover, UNet + + introduced a deep monitoring mechanism so that dense network structures can be pruned, which makes the model more flexible.

To further make up for the lost information, UNet3 + [27] redesigned the interconnections between encoder and decoder, In order to learn hierarchical representation from full-scale aggregated feature maps, full-scale in-depth supervision is further adopted in UNet3 + . UNet3 + produced accurate segmentation results, highlights organs and produces coherent boundaries. It is worth mentioning that it surpasses all previous state-of-the-art methods (PSPNet [11], DeepLab version [10–12], UNet + + [26], Attention UNet [28]) in quantitative testing on two datasets. Although UNet3 + has achieved excellent performance, there is still a lot of room for improvement.

In this paper, we present N-Net in order to obtain more accurate medical segmentation images. The main contributions of this paper are summarized as follows: (1) We propose N-Net to achieve accurate medical image segmentation. Unlike UNet + + and UNet3 + , we pay more attention to the extraction of feature in encoder and propose a dual encoder model. (2) We introduce the improved SE module [29] into the dual encoder model and evaluate the effectiveness of the SE module to improve the performance of the segmented network. (3) We propose a mixed loss function to better adapt to our experiment and evaluate the effectiveness of the mixed loss function to improve the segmentation accuracy. (4) We conduct abundant experiments on lung and liver datasets, where N-Net yields consistent improvements over a number of baselines.

2 Methods

2.1 Overview of N-Net architecture

Our N-Net structure is illustrated in Fig. 1, we improve the UNet architecture and propose the dual encoder model with the SE model. The convolutional layer connection is shaped like N, so we named this architecture N-Net. The two parallel paths of the dual encoder are connected layer by layer through ordinary skip connections. We let X_{DC}^{i}, X_{EN}^{i} and X_{DE}^{i} $(i = 1, 2, 3 \ldots)$, respectively, denote the convolutional layers of the three branches (two parallel branches of the dual encoder and the decoder branch) of N-Net. The structures of these two parallel convolutional layers of the dual encoder are similar, and they consist of two convolutional units. Each convolution unit consists of a convolution layer, a batch normalization layer and a ReLU activation layer. After each X_{DC}^{i} and X_{EN}^{i} $(i = 1, 2, 3 \ldots)$, the 2 × 2 max pooling reduces the size of the feature maps by half, In order to recover the lost spatial information in the pool layer, the decoder adopts a series of bilinear upsampling operations. After each upsampling operation, X_{DE}^{i} $(i = 1, 2, 3 \ldots)$, which consists of 320 filters of size 3 × 3, a batch normalization and a ReLU activation function, is appended. Finally, the probability map of segmentation is output by using 3 × 3 convolution layer and sigmoid activation function, and its size is the same as the original input.

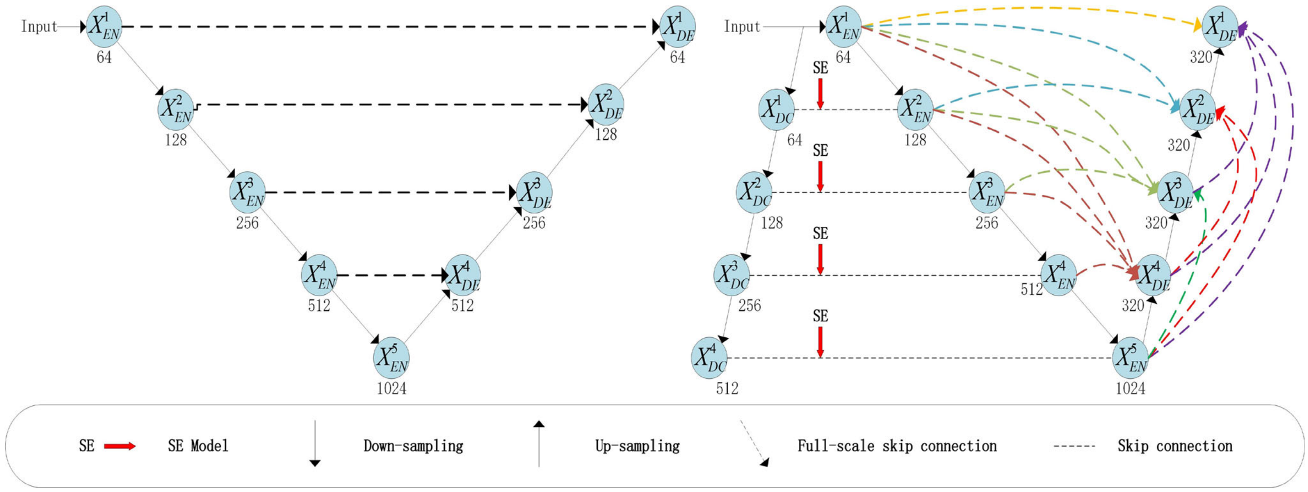

Fig. 1 UNet and N-Net architecture. The arrows represent the different operations

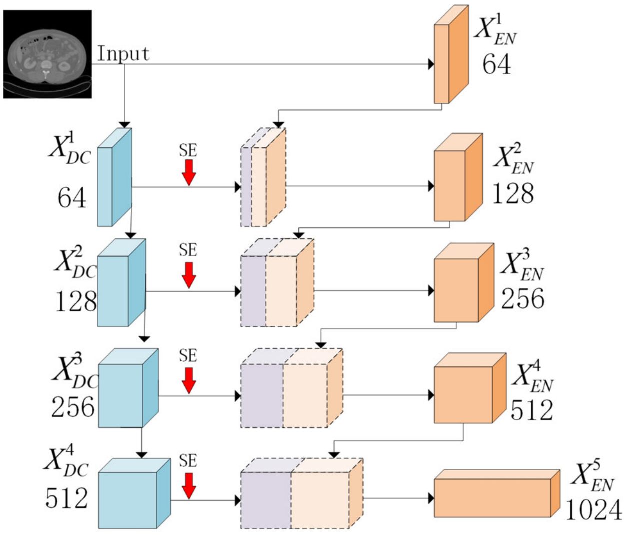

Fig. 2 The dual encoder model in N-Net

2.2 Dual encoder model

In the process of encoding and decoding, some information must be lost. In order to reduce the loss of information, we introduce a dual encoder model, which consists of X^i_{DC} $(i = 1, 2, 3, 4)$ and X^i_{EN} $(i = 1, 2, 3\ldots)$. As depicted in Fig. 2, the two parallel branches each have 4 max pooling layers with of size 2 × 2. Then, through the max pooling operation, the two parallel branches correspond to the same resolution of the convolutional layer. The difference between the X^i_{DC} and the X^i_{EN} $(i = 1, 2, 3\ldots)$ is that X^i_{DC} $(i = 1, 2, 3\ldots)$ use the dilated convolution to extract the feature information. Control the number of channels to allow the fusion of the feature maps of the X^i_{DC} and the X^i_{EN} $(i = 1, 2, 3\ldots)$ to provide the next layer of the X^{i+1}_{EN} $(i = 1, 2, 3\ldots)$ as input.

The dual encoder model not only deepen the network depth but also integrate comprehensive information. The dual encoder model and full-scale connections make up for the information lost by max pooling operations and integrate full-scale information to capture fine-grained details and coarse-grained semantics on a full-scale basis. The dual encoder model uses different convolutions to extract multi-scale feature information to enrich semantic information. The use of dilated convolution can obtain a larger receptive field. In the case of the same feature maps, the dilated convolution can improve the effect of small object recognition and segmentation in the task.

In order to further enhance the capability of N-Net feature extraction, we introduce the improved SE model in the skip connections of dual encoder model as Fig. 3. Each X^i_{DC} $(i = 1, 2, 3, 4)$ generates a side output from which the SE module learns the importance of different channel features. As Fig. 3 shows, the first step in SE model is the global average pooling (GAP) of the $H \times W$ features for each channel to get a scalar, which is called Squeeze. Given the input feature maps $F \in R^{C\times H\times W}$, GAP generates a feature vector $Z \in R^{C\times 1\times 1}$. GAP can be formally defined as:

$$Z_n = \frac{1}{H \times W} \sum_{x=0}^{H} \sum_{y=0}^{W} F_n(x, y) \tag{1}$$

where F_n is the nth channel feature map of F, $n \in \{1, 2\ldots, C\}$ and (x, y) is a pixel in F_n.

We employ two fully connected layers. The first fully connected layer reduces the number of channels to C/r, where r is the scaling factor determined by the results of subsequent experiments. And the number of channels backs to C after the second fully connected layers. The ReLU function after the first fully connected layer is employed to guarantee

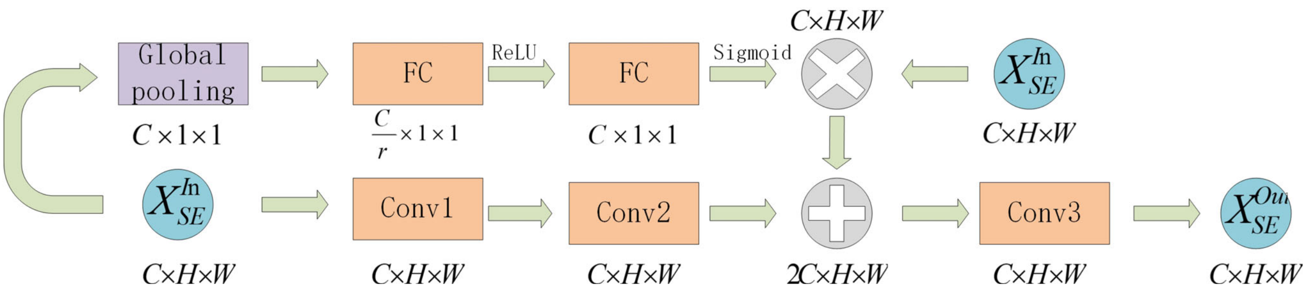

Fig. 3 An example to explain how the SE Model gets channel information

that multiple channels can be emphasized and the sigmoid function after the second one is performed to generate the non-linear relationships among different channels. The two FC get a weight value between 0 and 1. Finally, multiplicating each element of each $H \times W$ feature by the weight of the corresponding channel gives the new feature maps, which is called excitation. The weight value ω to recalibrate channels of input features is calculated as:

$$\omega = \sigma\left(W_2\delta\left(W_1\rho\left(X_{SE}^{In}\right)\right)\right) \tag{2}$$

where the $\sigma(\cdot)$ denotes the sigmoid function, and $\delta(\cdot)$ denotes the ReLU activation function. The $\rho(\cdot)$ denotes the global average pooling layer. Then, a dimensionality- reducing fully connected (FC) layer parameterized by $W_1 \in R^{\frac{C}{r}\times C}$ and a dimensionality-increasing FC layer parameterized by $W_2 \in R^{C\times\frac{C}{r}}$ are applied to obtain the "excitation" weight ω.

A complete SE block can be defined as a transformation mapping the input feature map $X_{\mathrm{SE}}^{\mathrm{In}} \in R^{C\times H\times W}$ to the output feature map $X_{\mathrm{SE}}^{\mathrm{Out}} \in R^{C\times H\times W}$:

$$X_{\mathrm{SE}}^{\mathrm{Out}} = H\left(F_{\mathrm{scale}}\left(\omega, X_{\mathrm{SE}}^{\mathrm{In}}\right) + H_1\left(X_{\mathrm{SE}}^{\mathrm{In}}\right)\right) \tag{3}$$

where transformation $\mathrm{H}_1(\cdot)$ represents two convolution operations with the batch normalization and the ReLU activation, and $\mathrm{H}(\cdot)$ represents a similar convolution operation to modify the dimension.$F_{\mathrm{scale}}\left(\omega, X_{\mathrm{SE}}^{\mathrm{In}}\right)$ refers to channel-wise multiplication between the weight ω and the input feature map $X_{\mathrm{SE}}^{\mathrm{In}}$.

This attention mechanism allows the model to pay more attention to the channel features with the most information, while suppressing those features that are not important. The whole operation can be regarded as learning the weight coefficients of each channel, thus making the model more discriminative to the characteristics of each channel. This can help us get more accurate information about the location and edges of the organs.

2.3 Loss function

Loss function is used to quantify the difference between the estimation of the network and the ground truth. The quality of the training model has a certain relationship with the loss function. Focal Loss [30] is a loss function commonly used in image detection and segmentation. Focal loss was proposed to address the extreme foreground–background class imbalance encountered during training of dense detectors in the object detection mission, which is defined as

$$L_{\mathrm{FL}}(\rho_t) = -\alpha_t(1-\rho_t)^{\gamma}\log(\rho_t) \tag{4}$$

$$\rho_t\begin{cases} p & \text{if } y=1 \\ 1-p & \text{otherwise}\end{cases} \tag{5}$$

$$\alpha_t\begin{cases} \alpha & \text{if } y=1 \\ 1-\alpha & \text{otherwise}\end{cases} \tag{6}$$

where $p \in [0, 1]$ is the model's estimated probability for a class labeled $y = 1$, $\alpha \in [0, 1]$ is a weighting factor, and $\gamma \geq 0$ is a tunable focusing parameter. Focal loss sets $(1-\rho_t)^{\gamma}$ as modulating factor. The aim is to reduce the weight of the easily classified samples so that the model can focus more on the difficult classified samples during training. The parameters of focal loss are set to $\gamma = 2$, $\alpha = 0.25$ as per [30].

Luo et al. introduce a novel soft IoU loss [31] to obtain more accurate road topology in aerial images. This loss function is differentiable and thus amenable to back propagation. It was defined by replacing the indicator functions with the softmax outputs, and it could be defined in as follows:

$$L_{\mathrm{IoU}} = 1 - \frac{1}{2}\left(\frac{\sum_i p_{i1} * p_{i1}^*}{\sum_i p_{i1} + p_{i1}^* - p_{i1} * p_{i1}^*} + \frac{\sum_i p_{i0} * p_{i0}^*}{\sum_i p_{i0} + p_{i0}^* - p_{i0} * p_{i0}^*}\right) \tag{7}$$

where p_{ix} is the prediction score at location i for class x, and p_{ix}^* is the ground truth distribution which is a delta function at y_i^*, the correct label.

In order to capture the best characteristics of both loss functions and train our task more easily and obtain more accurate segmentation effect, we propose to combine them:

$$\text{loss}_{\text{mix}} = \beta * \text{loss}_{\text{IoU}} + (1 - \beta) * \text{loss}_{\text{FL}} \tag{8}$$

β is the parameter in the equation. According to a large number of experiments, we constantly adjust the β to compare the experimental results, and we empirically set $\beta = 0.15$.

3 Experiments and discussion

3.1 Datasets

The datasets in our experiment are obtained through two different competitions. The first dataset is the lung dataset in the Kaggle. This dataset about lung includes 267 CT images and the corresponding 267 mask images labeled by experts. We also add the liver dataset in the Chaos2019 [32–34] to prove the universality of the model. We select the CT images of 20 different patients. In total, 2377 slices will be provided for training. Finally, we use 3616 COVID-19 disease X-ray from version 4 of the COVID-19 Radiography Database [35] to prove the universality of our model. Both the datasets consist of RGB images containing different regions-of-interest (ROI) with image size of 512 × 512 pixels. Each of them is divided into three subsets, where 70% of the images are used for training, 10% for validation, and 20% for testing. The network is trained for 100 epochs. In order to save training time, the datasets are resampled to 240 × 240 pixels for our experiments. In addition, the training sets are additionally expanded by data augmentation methods including rotation, translation, and flip.

3.2 Training

The proposed N-Net was trained on a NVIDIA RTX 2060 SUPER GPU, with 12 GB of RAM. The implementation of our method is based on Python 3.6 under the same conditions of a personal computer equipped with PyTorch framework. We utilize the stochastic gradient descent to optimize our network and its hyper-parameters are set to the default values. The loss function used in all comparison models is Focal loss. Dice coefficient is used as the main evaluation metric for each case.

3.3 Experiments

3.3.1 Evaluation metrics

In order to evaluate our method quantitatively, we use four evaluation metrics including Dice, Recall, Precision and F1 score. Among them, the four evaluation indicators can be defined as

$$\begin{aligned}
\text{Dice} &= \frac{2 \times \text{TP}}{(\text{TP} + \text{FN}) + (\text{TP} + \text{FP})}, \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
F1 &= \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}
\end{aligned} \tag{9}$$

where TP is the number of pixels with label 1 and predicted value 1, TN is the number of pixels with label 0 and predicted value 0, FP is the number of pixels with label 0 and predicted value 1, and FN is the number of pixels with label 1 and predicted value 0.

For these metrics, the higher the value, the better the segmentation effect. The high Recall means that the more true mask pixels are predicted. The high Precision indicates that the true value of the mask accounts for a large proportion of the prediction results. The high Dice coefficient indicates that the prediction result is highly similar to the true value, and the model is excellent. F1 score can be regarded as a harmonic average of the Precision and Recall of the model.

3.3.2 Comparison with UNet, UNet + + , UNet3 +

We compare the proposed N-Net with UNet, UNet + + , UNet3 + and use Focal loss as the loss function to train the network to ensure fair comparison. Tables 1, 2 and 3 compare the number of Dice, Recall, Precision and F1 score of UNet, UNet + + , UNet3 + and the proposed N-Net architecture on the datasets. In Table 4, we also recorded the time spent

Table 1 Comparison of U-Net, U-Net + + , U-Net3 + and N-Net on lung dataset

Architecture	Dice	Recall	Precision	F1
UNet	0.9203	0.8750	0.9733	0.9215
UNet + +	0.9218	0.8723	0.9844	0.9249
UNet3 +	0.9246	0.8798	0.9815	0.9278
N-Net	**0.9329**	**0.8950**	**0.9856**	**0.9381**

Table 2 Comparison of U-Net, U-Net + + , U-Net3 + and N-Net on liver dataset

Architecture	Dice	Recall	Precision	F1
UNet	0.8154	0.7614	0.9302	0.8374
UNet + +	0.8277	0.7927	0.9137	0.8489
UNet3 +	0.8349	0.7820	0.9398	0.8537
N-Net	**0.8619**	**0.8118**	**0.9508**	**0.8758**

Table 3 Comparison of U-Net, U-Net + + , U-Net3 + and N-Net on COVID-19 dataset. The best results are highlighted in bold

Architecture	Dice	Recall	Precision	F1
UNet	0.9455	0.9079	0.9834	0.9441
UNet + +	0.9489	0.9112	0.9872	0.9476
UNet3 +	0.9525	0.9237	0.9853	0.9535
N-Net	**0.9559**	**0.9245**	**0.9892**	**0.9557**

Table 4 Time spent of N-Net and other 3 precious models

Method	Train time (s)	Test time (s)
UNet	**635**	1.567
UNet + +	1500	1.321
UNet3 +	2609	**1.042**
N-Net	2354	1.301

training 40 epochs and the time spent testing five images on lung dataset. It is worth mentioning that all results are directly from single-model test without relying on any post-processing tools. As seen, the segmentation performance of UNet + + and UNet3 + on the datasets is better than UNet. This improvement is attributed to both models improve the ordinary skip connections of UNet. N-Net achieves a excellent performance gain over both UNet + + and UNet3 + , obtaining average improvement of 0.023, 0.018 and 0.0104 point in Dice coefficient. N-Net not only has a good performance in the Dice coefficient, but also has the highest Recall rate among the four models compared. Its Recall is 0.01 to 0.03 points higher than that of the other three models, indicating that N-Net can be closer to the ground truth and obtain more accurate segmentation results. Moreover, the model proposed by us has the highest Precision, indicating that the predicted values of the proposed model have more effective values and there are less redundant error pixels.

To visualize the impact of the different encoder–decoder models, Figs. 4, 5 and 6 display the segmentation results on six images on the lung, liver and COVID-19 datasets. The proposed N-Net network achieves qualitatively better results than other encoder–decoder networks. It can be observed that our proposed method not only accurately localizes organs but also produces coherent boundaries, even in small object circumstances. Compared with our proposed model, the results

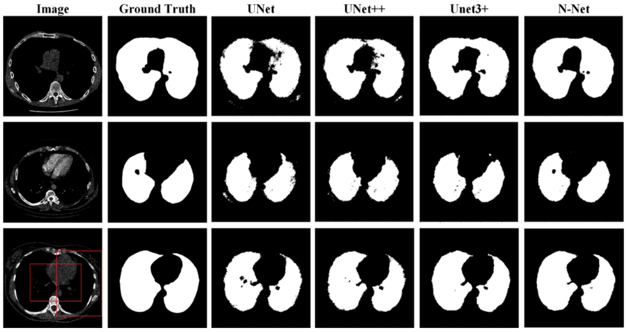

Fig. 4 Results on three images on the lung dataset

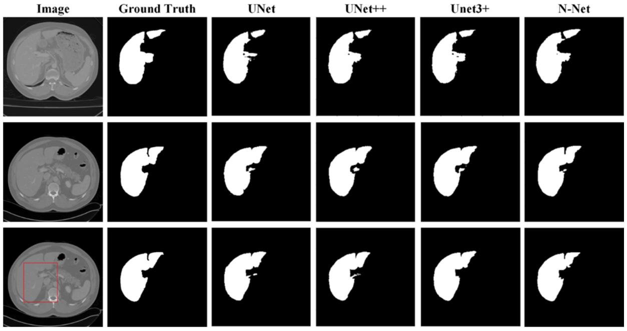

Fig. 5 Results on three images on the liver dataset

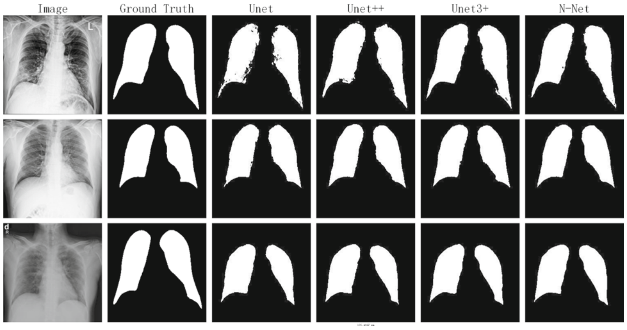

Fig. 6 Results on three images on the COVID-19 dataset

Table 5 Quantitative comparison among different values of r

r	Dice	Recall	Precision	F1	Params (M)
2	0.9350	0.9004	0.9793	0.9377	41.37
4	0.9360	0.9027	0.9788	0.9392	41.19
8	**0.9377**	**0.9053**	**0.9794**	**0.9408**	**41.11**
16	0.9357	0.9015	0.9794	0.9388	41.06

Table 6 Comparison of N-Net and other 3 precious models

Method	$\text{Dice}_{\text{Lung}}$	$\text{Dice}_{\text{Liver}}$	$\text{Dice}_{\text{Covid}}$
PSPNet[8]	0.9258	0.8247	0.9512
DeepLabV3[10]	0.9215	0.8233	0.9475
CE-Net[36]	0.9261	0.8337	0.9537
N-Net	**0.9329**	**0.8619**	**0.9559**

Table 7 Gain comparison after N-Net uses SE module and the mixed loss

Architecture	$\text{Dice}_{\text{Lung}}$	$\text{Dice}_{\text{Liver}}$	$\text{Dice}_{\text{Covid}}$
N-Net(focal loss)	0.9329	0.8619	0.9559
N-Net(SE Model + focal loss)	0.9377	0.8670	0.9592
N-Net(SE Model + UNet loss[13])	0.9382	0.8667	0.9590
N-Net(SE Model + mixed loss)	**0.9412**	**0.8703**	**0.9633**

of other three models have more obvious organ confusion. In the three figures, we can see that in the segmentation results of UNet and UNet + + , the boundary of the organ is relatively blurred, and it even breaks where the organ joins together. Both UNet3 + and N-Net are able to produce smooth lung segmentation edges. But N-Net segments the organ margins in greater detail. These visual results indicate that our approach can successfully recover finer segmentation details.

3.3.3 Estimation of r of SE model

As described in Sect. 2.2, the scaling factor r is a hyperparameter that allows us to vary the capacity and computational cost of the SE module, which needs to be set carefully. To investigate the tradeoff between the performance determined by r and the computational burden, we experiment with a range of different values of r.

As shown in Table 5, the larger the r value, the greater the extrusion of features, and the fewer parameters (Params) of N-Net are needed. Note that when $r = 8$, the values of Dice, Recall, Precision, and F1 reach the maximum values. However, the number of parameters increased by 0.05 M. Therefore, we set $r = 8$ to trade off performance against computational burden.

3.3.4 Comparison with other previous models

To further verify the superiority of our network, we compare N-Net with other previous models in this section. A quantitative evaluation of the results is presented in Table 6, from which we can observe that the proposed method is competitive with other existing methods by achieving the best values. And we can conclude that the proposed N-Net achieves the state-of-the-art results.

3.3.5 Ablation experiments

To demonstrate the effectiveness of the mixed loss function proposed in this method, we briefly compare the segmentation accuracy between N-Net with the mixed loss and that with focal loss and UNet loss. The quantitative results are given in Table 7, from which we can observe that the mixed loss outperforms the work of focal loss and UNet loss. Hence, we can conclude that mixed loss is more appropriate for medical image segmentation.

In order to further demonstrate the improvement effect of introducing SE module on the dual encoder model, we conduct ablation experiments. Table 7 summarizes the quantitative comparison results. Moreover, taking advantages of the SE module, the dual encoder model of N-Net learns the importance of different channel features.

4 Discussion

We enlarge the position marked by the red boxes in Figs. 4 and 5 to fully display the details of segmentation results. The artifacts and noises in CT images often affect the segmentation results of models. As shown in Fig. 7, there are obvious holes and spots in the segmentation results compared with the ground truth. These holes and spots show that models do not effectively distinguish artifacts and noises, which leads to the production of some false negatives and false positives. It can be seen that our proposed model can effectively reduce the generation of such false negatives and false positives. This is because the models lack sufficient information

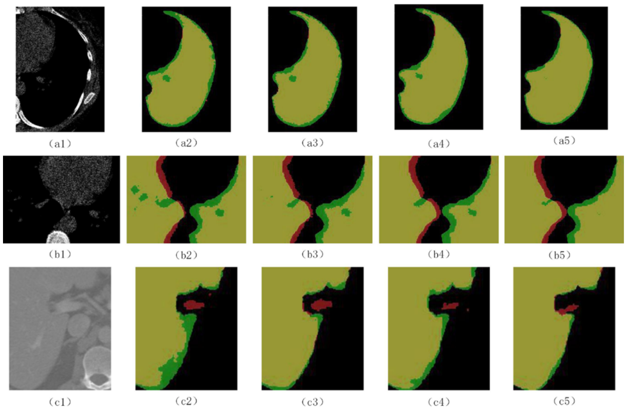

Fig. 7 Illustration of details of segmentation results. (a1, b1, c1) are enlarged CT images. (a2)–(c2), (a3)–(c3), (a4)–(c4) and (a5)–(c5) are the segmentation results of UNet, UNet + + , UNet3 + and N-Net, respectively. The yellow part indicates the correct segmentation area, and the red part and the green part indicate the false positive area and the false negative area, respectively

to distinguish artifacts and noises. Compared with the other three models, the dual encoder model of N-Net can extract features more effectively and obtain more comprehensive information. Before generating the final segmentation mask, the richer feature maps provided by the dual encoder model help the decoder obtain more image information. The above experiments demonstrate that N-Net can effectively reduce the false positives and false negatives of the segmentation results.

5 Conclusion

In this paper, we propose a medical image segmentation method based on the improved UNet model, called N-Net. Based on the encoder and decoder structure of UNet, a novel dual encoder model is proposed to aggregate more context features. The SE module and the mixed loss function are further introduced to yielding more accurate segmentation. Extensive experiments are conducted to assess the impact of the proposed model. We also compare our model with advanced encoder–decoder structure networks. Experimental results show that the proposed model is superior to the compared methods both quantitatively and qualitatively. This proves the efficiency of our method in providing accurate and reliable automatic segmentation of medical images.

Acknowledgements The authors thank the editors and the anonymous reviewers for their constructive and helpful comments.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A., Sánchez, C.I.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
2. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. Proc. Adv. Neural. Inf. Process. Syst. **25**, 1090–1098 (2012)
3. Szegedy, C., Liu, W. et al.: Going deeper with convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–92. (2015)
4. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. (2016)
6. Sun, X., Wu, P., Hoi, S.: Face detection using deep learning: an improved faster RCNN approach. Neurocomputing **299**, 42–50 (2018)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. (2015)
8. Zhao, H.S., Shi, J.P., Qi, X.J., Wang, X.G., Jia, J.Y.: Pyramid scene parsing network. In: The IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890. (2017)
9. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934. (2017)
10. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587. (2017)
11. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 834–848 (2018)
12. Chen, L.-C., Zhu, Y.K., Papandreou, G., Adam, H.: Encoder - decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, arXiv preprint arXiv:1802.02611 (2018)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. (2015)
14. Tong, N., Gou, S., Yang, S., Ruan, D., Sheng, K.: Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. Med. Phys. **45**(10), 4558–4567 (2018)
15. Wang, Y., Gong, G.Z., Kong, D.T., Li, Q., Dai, J.P., Zhang, H.Y., Qu, J.H., Liu, X.Y., Xue, J.: Pancreas segmentation using a dual-input v-mesh network. Med. Image Anal. **69**, 101958 (2021)
16. Xue, J., Wang, Y., Kong, D.T., Wu, F.Y., Yin, A.J., Qu, J.H., Liu, X.Y.: Deep hybrid neural-like P systems for multiorgan segmentation in head and neck CT/MR images. Expert Syst. Appl. **168**, 114446 (2021)
17. Hesamian, M.H., Jia, W., He, X.J., Kennedy, P.: Deep learning techniques for medical image segmentation: achievements and challenges. J. Dig. Imag. **32**, 582–596 (2019)
18. Liu, X., Song, L., Liu, S., Zhang, Y.: A review of deep-learning-based medical image segmentation methods. Sustainability **13**(3), 1224 (2021). https://doi.org/10.3390/su13031224

19. 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)
20. Jin, Q., Meng, Z., Sun, C., Wei, L., Su, R.: RA-UNet: a hybrid deep attention-aware network to extract liver and tumor in CT scans. Front. Bioeng. Biotechnol. **8**, 1471 (2020)
21. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imag. **37**, 2663–2674 (2018)
22. Adiga V,S., Sivaswamy, J.: FPD-M-net: fingerprint image denoising and inpainting using m-net based convolutional neural networks. In: Painting and Denoising Challenges (2019).
23. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. (2016).
24. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ben Ayed, I.: HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans. Med. Imag. **38**, 1116–1126 (2018)
25. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imag. **37**, 2514–2525 (2018)
26. Zhou, Z.W., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.M.: UNet++: a nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. (2018)
27. Huang, H., Lin, L., Tong, R., Hu, H., Wu, J.: UNet 3+: a full-scale connected UNet for medical image segmentation. arXiv preprint arXiv:2004.08790 (2020)
28. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M.: Attention U-Net: learning where to look for the pancreas. (2018)
29. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 2011–2023 (2020)
30. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In IEEE International Conference on Computer Vision, pp. 2999–3007. (2017)
31. Máttyus, G., Luo, W., Urtasun, R.: DeepRoadMapper: extracting road topology from aerial images. In IEEE International Conference on Computer Vision, pp. 3458–3466. (2017)
32. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., et al.: CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Med. Image Anal. **69**, 101950 (2021)
33. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: CHAOS - combined (CT-MR) healthy abdominal organ segmentation challenge data (Version v1.03), (2019)
34. Kavur, A.E., Gezer, N.S., Barış, M., Şahin, Y., Özkan, S., Baydar, B., et al.: Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors. Diagn. Interv. Radiol. **26**, 11–21 (2020)
35. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Chowdhury, M.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-rays images. Comput. Biol. Med. **132**, 104319 (2021)
36. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: CE-Net: context encoder network for 2D medical image segmentation. IEEE Trans. Med. Imag. **38**, 2281–2292 (2019)

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