2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230609

CMU-NET: A STRONG CONVMIXER-BASED MEDICAL ULTRASOUND IMAGE SEGMENTATION NETWORK

Fenghe Tang[1], *Lingtao Wang*[1], *Chunping Ning*[2], *Min Xian*[3], *and Jianrui Ding*[1⋆]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.
[2]Ultrasound Department, The Affiliated Hospital of Qingdao University, Qingdao, China.
[3]Department of Computer Science, University of Idaho, Idaho Falls, ID 83401, USA

ABSTRACT

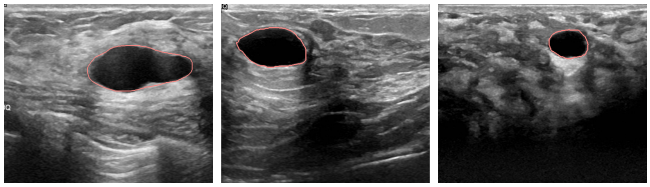
U-Net and its extensions have achieved great success in medical image segmentation. However, due to the inherent local characteristics of ordinary convolution operations, U-Net encoder cannot effectively extract global context information. In addition, simple skip connections cannot capture salient features. In this work, we propose a fully convolutional segmentation network (CMU-Net) which incorporates hybrid convolutions and multi-scale attention gate. The ConvMixer module extracts global context information by mixing features at distant spatial locations. Moreover, the multi-scale attention gate emphasizes valuable features and achieves efficient skip connections. We evaluate the proposed method using both breast ultrasound datasets and a thyroid ultrasound image dataset; and CMU-Net achieves average Intersection over Union (IoU) values of 73.27% and 84.75%, and F1 scores of 84.16% and 91.71%. The code is available at https://github.com/FengheTan9/CMU-Net.

Index Terms— Ultrasound image segmentation, U-Net, ConvMixer, multi-scale attention

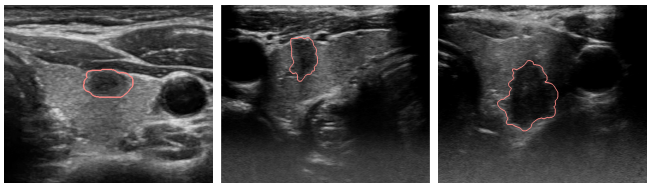
1. INTRODUCTION

Ultrasound imaging is non-invasive, non-radiative, cost effective and real-time for disease detection. It has been widely used in the detection of breast tumor, thyroid nodule, fetal, and gonadal tissue [1]. Conventional disease detection using ultrasound images depended on manual labeling, which is laborious and time-consuming. The results were sensitive to subjective factors such as radiologists' experience and mental state. With the emergence of deep learning approaches, automatic medical image segmentation has been rapidly developed in the field of image analysis, which can effectively overcome the above limitations.

The segmentation of medical ultrasound images is challenging. As shown in Fig.1, most ultrasound images only contain one lesion, and binary segmentation approaches could be applied. But the sizes, shapes, and texture patterns of lesions from different cases vary greatly. In addition, ultrasound images usually have low contrast, high speckle noises, and shadow artifacts, and conventional segmentation approaches used to perform poorly.



(a) Breast ultrasound image



(b) Thyroid ultrasound image

Fig. 1. Examples of ultrasound image segmentation. The pink contours denote lesion boundaries.

U-Net [2] has an encoder-decoder based segmentation architecture. It can effectively fit scarce medical image data. In recent years, many medical segmentation networks based on U-Net have been proposed, such as U-Net++ [3], Attention U-Net [4], Unet3+ [5], and UNeXt [6]. Due to the locality of ordinary convolution operations in U-Net, a number of networks based on Transformer [7] have recently been applied to medical image segmentation tasks [8, 9, 10] to extract global information of images. TransUnet [8] employed Vit [11] to obtain global context with CNN, but it required massive medical images and computing overhead.

In order to solve the limitation of ordinary convolution locality, Trockman et al. proposed the ConvMixer [12] which used large convolutional kernels to mix remote spatial locations to obtain global context information. Compared with the Transformer, the ConvMixer is more efficiency and adapt to computer vision tasks better, and its computational overhead is less than that of the self-attention mechanism.

Inspired by the U-shape architectural design and ConvMixer, we propose an efficient fully convolutional image segmentation network, namely CMU-Net, which contains the

ConvMixer module and multiscale attention gate. The ConvMixer module is used to extract global context information. The multi-scale attention gate suppresses irrelevant features and strengthen the valuable features.

This work makes the following contributions: 1) We propose a strong fully convolution medical image segmentation network based on ConvMixer; 2) the proposed multi-scale attention gates effectively transfer knowledge using skip-connection; and 3) we successfully improve the performance of medical image segmentation using breast and thyroid ultrasound images.

2. METHOD

The network architecture of the proposed CMU-Net is shown in Fig.2. The CMU-Net is divided into the encoder and decoder stages with skip-connection. In the encoder stage, high-level semantic information of medical images is extracted through ordinary convolutions, and the feature maps are input into the ConvMixer module to obtain mixed spatial and location information. In the decoder stage, the features from multi-scale attention gates are spliced with the corresponding up-sampling features to achieve accurate positioning.

2.1. Encoder stage

As shown in Fig.2, the encoder has five levels of convolutions from top to bottom. Each level consists of two ordinary convolution blocks and a down sampling operation. Specifically, each ordinary convolution block is equipped with a convolution layer, a batch normalization layer and ReLU activation. The kernel size is 3×3, stride of 1 and padding of 1. The down sampling of the encoder is max pooling with window size of 2×2. At the last level, the feature map is input into the ConvMixer block [12] which is composed of L ConvMixer layers. A single ConvMixer layer consists of depthwise convolution (i.e., kernel size k×k) and pointwise convolution (i.e., kernel size 1×1). The number of group channels of the depthwise convolution kernel is equal to the channels of the input feature map. Each convolution is with a GELU [13] activation and a batch normalization, and is defined by

$$f_l' = BN(\sigma_1\{DepthwiseConv(f_{l-1})\}) + f_{l-1} \quad (1)$$

$$f_l = BN(\sigma_1\{PointwiseConv(f_l')\}) \quad (2)$$

Where f_l represents the output feature map of layer l in the ConvMixer block, σ_1 represents the GELU activation, and BN represents the batch normalization. Since the feature maps from all layers in the ConvMixer module maintain the same resolution and size, we directly up-sample the features extracted by the ConvMixer block.

2.2. Decoder stage with skip connection

The decoder also has five modules from bottom to top. Each module is composed of two ordinary convolution blocks and an upsampling block. Specifically, the upsampling block is equipped with an upsampling layer, a convolution layer, a batch normalization layer and ReLU activation. The bilinear interpolation is utilized to upsample the feature maps. The convolution kernels have size of 3×3, stride of 1 and padding of 1.

We propose the multi-scale attention gates and integrate it with the skip connections. It is used to suppress unimportant features and enhance the valuable features. Specifically, the implementation of the multi-scale attention gate is shown in Fig.3.

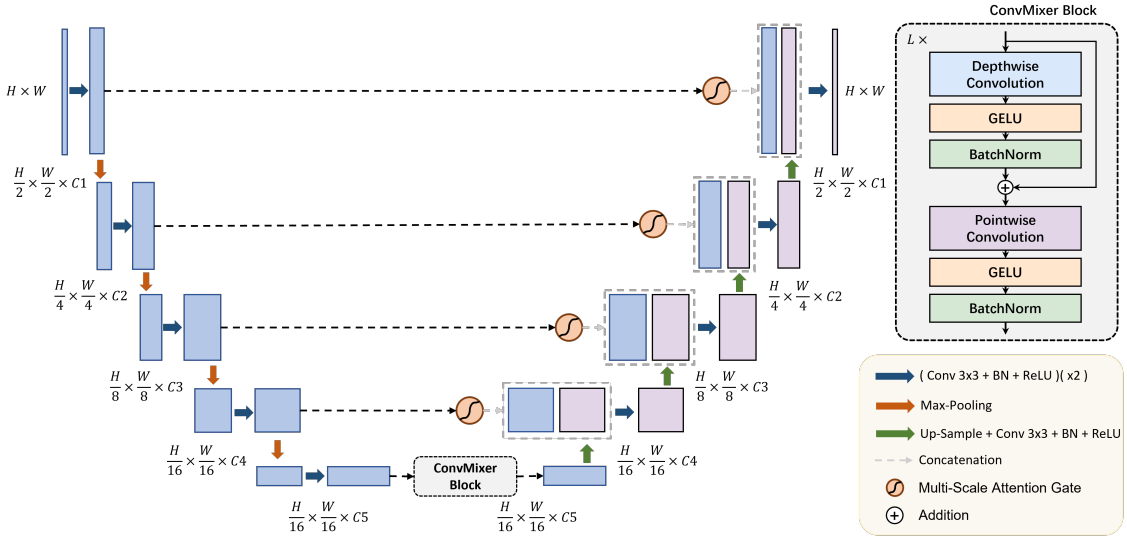


Fig. 2. Overview of the proposed CMU-Net architecture. Note that the channel numbers are adopted from U-Net, i.e., C1 = 64, C2 = 128, C3 = 256, C4 = 512, and C5 = 1024.

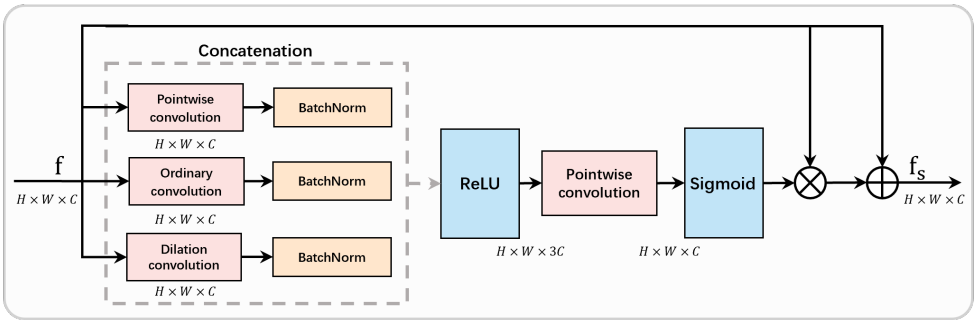


Fig. 3. Multi-scale attention gate.

To select features according to different resolutions adaptively, we develop three convolutions with different receptive fields to extract features: pointwise convolution, ordinary convolution (i.e., kernel size of 3×3 and stride of 1 and padding of 1) and dilated convolution (i.e., kernel size of 3×3, stride of 1, padding of 2 and dilation rate of 2). Each convolution is with a batch normalization layer. The three different convolutions generate feature maps with same sizes; and we concatenate the output feature maps before a ReLU activation and vote to select valuable features by another pointwise convolution:

$$f_{Concat} = \sigma_2(Concat\{BN\{PointwiseConv(f)\}, BN\{OrdinaryConv(f)\}, BN\{DilationConv(f)\}\}) \quad (3)$$

$$f_s = f \times \sigma_3(PointwiseConv(f_{Concat})) + f \quad (4)$$

Where f represents encoding features, f_{Concat} is the concatenated feature, f_s is the output feature from the multi-scale attention gate, and σ_2 and σ_3 denotes the ReLU and Sigmoid activation, respectively.

3. RESULTS AND DISCUSSION

3.1. Datasets, evaluation and implementation details

The Breast UltraSound Images (BUSI) [14] and private Thyroid UltraSound dataset (TUS) are utilized to evaluate the proposed approach. BUSI collected 780 breast ultrasound images, including normal, benign and malignant cases of breast cancer with their corresponding segmentation results. We only use benign and malignant images (647 images). TUS is collected from the Ultrasound Department of the Affiliated Hospital of Qingdao University. It includes 192 cases, totally 1,942 images with segmentation results from three experienced radiologists. We adopt five commonly used metrics to quantitatively evaluate the performance of different segmentation models: Intersection over Union (IoU), Recall, Precision, F1-value and Accuracy.

The loss L between the predicted map $\hat{y}$ and ground truth target map y is defined as a combination of binary cross entropy (BCE) and dice loss (Dice),

$$L = 0.5BCE(\hat{y}, y) + Dice(\hat{y}, y) \quad (5)$$

The experiments use Adam optimizer to optimize the network. The initial learning rate is set to 0.0001, and momentum is 0.9. The batch size is set to 8, and the number of epochs is 300. The two datasets are randomly split thrice, 80% for training and 20% for validation. In addition, we resize all images to 256×256 and perform random rotation and flip for data augmentation.

3.2. Results

We compare CMU-Net, U-Net [2], U-Net++ [3], Attention U-Net [4], U-Net3+ [5], TransUnet [8] and UNeXt [6]. Note that the encoder of U-Net++ is ResNet34 [15]. Moreover, we conduct experiments with different ConvMixer depths and

Table 1. Results on the BUSI dataset (%).

	IoU	Recall	Precision	F1-value	Accuracy
U-Net[2]	68.49±0.18	80.57±2.24	82.52±2.34	80.88±0.07	96.74±0.08
Attention U-Net[4]	70.38±1.48	81.44±1.67	83.66±0.61	82.16±0.97	96.99±0.12
U-Net++[3]	69.49±0.15	81.27±1.36	81.87±1.07	81.15±1.25	96.34±0.22
U-Net3+[5]	65.39±0.12	77.54±2.02	80.66±2.50	78.22±0.07	95.96±0.09
TransUnet[8]	66.75±1.50	78.65±4.32	81.33±2.71	79.46±1.05	96.24±0.38
UNeXt[6]	66.76±0.05	77.25±1.43	83.49±1.28	79.72±0.12	96.60±0.05
CMU-Net	**73.27±0.43**	**84.26±0.54**	**84.81±1.32**	**84.16±0.47**	**97.33±0.14**

Table 2. Results on TUS dataset (%).

	IoU	Recall	Precision	F1-value	Accuracy
U-Net[2]	83.51±0.10	90.15±0.91	92.00±0.83	90.97±0.05	99.21±0.01
Attention U-Net[4]	83.90±0.14	90.87±0.58	91.71±0.41	91.21±0.08	99.22±0.01
U-Net++[3]	84.23±0.33	90.59±0.35	92.34±0.32	91.40±0.20	99.22±0.03
U-Net3+[5]	83.60±0.14	90.21±0.90	92.02±0.78	91.01±0.07	99.18±0.01
TransUnet[8]	82.75±0.25	89.51±0.19	91.66±0.23	90.47±0.13	99.13±0.02
UNeXt[6]	81.19±0.18	88.41±1.13	90.86±1.17	89.50±0.07	99.05±0.02
CMU-Net	**84.75±0.30**	**91.53±0.37**	**92.02±0.13**	**91.71±0.17**	**99.27±0.01**

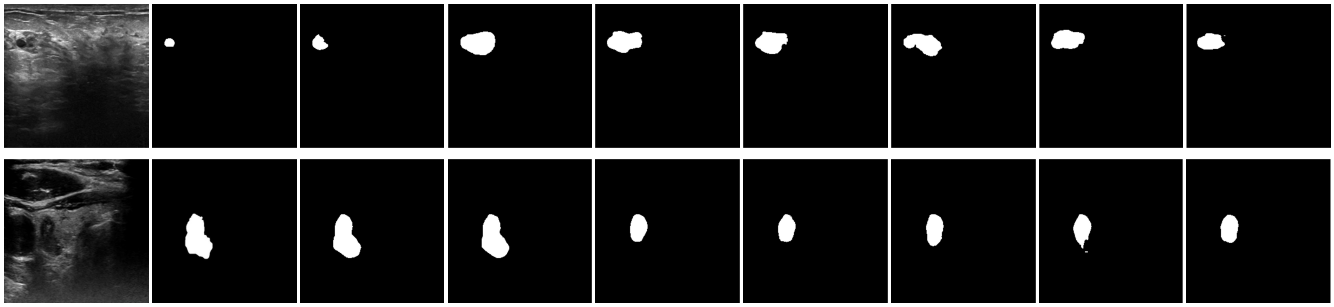
kernel sizes and finally the ConvMixer block with a depth of 7 and a kernel size of 7 can achieve the best performance.

As shown in Tables 1 and 2, CMU-Net obtains better segmentation performance than all other six approaches. Especially, on the BUSI dataset, CMU-Net achieves the highest performance and improves by 2.89% in IoU and 2.00% in F1 score. Further for the TUS dataset, CMU-Net improves all metrics compared with other models and achieves a better trade-off between recall and specificity. The TransUnet, which consists of hybrid CNN and Transformer structures, needs to feed a large amount of training data, performs unsatisfactorily on small datasets. On the contrary, CMU-Net demonstrates its advantages with consideration to both local and global information, and highlights features that are meaningful for the task. Some sample results are shown in Fig.4. It can be seen that CMU-Net generates more accurate lesion regions and shapes.

Moreover, we conducted an ablation study to analyze the contribution of each module in the CMU-Net. As shown in Table 3, we can see that the performance of the original U-Net has been greatly improved after the introduction of the ConvMixer block. Next, we add multi-scale attention gates to improve the performance further, and it shows that the multi-scale attention gates can effectively magnify the affection of helpful encoder features in knowledge transfer.

4. CONCLUSION AND PERSPECTIVES

In this work, we propose the CMU-Net, a fully convolutional network for medical ultrasound segmentation. We introduce ConvMixer block into a U-shape network architecture to build a strong encoder for obtaining global context information, and propose a multi-scale attention gate module for emphasizing valuable features to achieve efficient skip connections. We validate CMU-Net on two ultrasound datasets, and it achieves the state-of-the-art-performance. In the future, more experiments can be carried out on CMU-Net, such as using larger convolution kernels, placing ConvMixer blocks at different encoder levels. Further analyzing errors to improve the accuracy. Combining with the physiological and anatomical structures of the lesions to improve the interpretability of the model.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 4. Segmentation result. Row 1 - BUSI dataset, Row 2 – TUS dataset. (a) Input and (b) Ground Truth. Predictions of (c) CMU-Net, (d) Attention U-Net, (e) TransUnet, (f) U-Net, (g) U-Net++ (h), U-Net3+, and (i) UNeXt.

Table 3. Ablation study on the BUSI dataset (%).

	IoU	F1-value	Accuracy
Original U-Net	68.49±0.18	80.88±0.07	96.74±0.08
U-Net + ConvMixer	72.36±0.37	83.57±0.30	97.23±0.04
U-Net + ConvMixer + Multi-scale attention gate	73.27±0.43	84.16±0.47	97.33±0.14

5. COMPLIANCE WITH ETHICAL STANDARDS

Informed consent was obtained from all individual participants involved in the study.

6. ACKNOWLEDGMENTS

This work is supported by Shandong Natural Science Foundation of China (ZR2020MH290) and by the Joint Funds of the National Natural Science Foundation of China (U22A2033).

7. REFERENCES

[1] Qinghua Huang, Fan Zhang, and Xuelong Li, "Machine learning in ultrasound computer-aided diagnostic systems: a survey," *BioMed research international*, vol. 2018, 2018.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "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. Springer, 2018.

[4] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[5] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[6] Jeya Maria Jose Valanarasu and Vishal M Patel, "Unext: Mlp-based rapid medical image segmentation network," *arXiv preprint arXiv:2203.04967*, 2022.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[9] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 36–46.

[10] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 109–119.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] Asher Trockman and J Zico Kolter, "Patches are all you need?," *arXiv preprint arXiv:2201.09792*, 2022.

[13] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[14] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, pp. 104863, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.