Dean&Francis
ISSN 2959-6157

CMUeXt network model: Analysis and Verification of Insufficient Segmentation Performance on CT and MRI Images

Dayou Yang

School of Engineering, University
of Bristol, Bristol, UK, BS81QU
571410977@qq.com

Abstract:

Multimodal medical image segmentation is a critical auxiliary technique for clinical diagnosis, but it still confronts notable challenges in addressing noise interference, boundary blurring, and complex anatomical structures in CT, MRI, and other images. This article systematically evaluates the practical performance of the lightweight network CMUeXt, which integrates large convolution kernel and skip fusion module in CT and MRI segmentation tasks. Experiments are conducted on an NVIDIA RTX 4090 GPU using public datasets: CT (lung/liver), MRI (brain tumor), and mixed-modal datasets. The evaluation adopts Intersection over Union (IoU) and F1-score as core metrics, supplemented by visual validation. Results demonstrate significant limitations of CMUeXt: weak boundary perception ability, ambiguous probability maps, and poor cross-modal generalization. Its large-kernel design and skip fusion module fail to adapt to inherent differences in multimodal data. Future optimizations should integrate dynamic convolution, attention mechanisms, and domain adaptation to enhance its practical clinical applicability.

Keywords: Medical image segmentation, lightweight network, large kernel convolution, CMUNeXt, performance verification, Model Limitations

1. Introduction

Medical imaging, particularly CT and MRI, plays a pivotal role in modern clinical practice. CT excels in visualizing bony structures, while MRI offers superior soft tissue contrast. Deep learning has advanced medical image segmentation, with models like U-Net[1] and the Visual Transformer(ViT)[2] achieving notable performance. However, their high computational complexity often hinders deployment in resource-constrained clinical environment. This has spurred interest in lightweight architectures like CMUeXt[3], which leverages large convolution kernels for global context modeling and aims to balance segmentation accuracy with computational efficiency on challenging multimodal CT/MRI data..

This article seeks to explore the practical performance and limitations of the CMUNeXt network in multimodal medical image segmentation tasks (CT and MRI) through systematic experimental validation. The research focuses on three core objectives: (1) Evaluate the segmentation accuracy and robustness across CT and MRI tasks; (2) Analyze the adaptability deficiencies of its core modules (large convolution kernel, Skip Fusion) under different modal data; (3) Reveal common limitations of lightweight design in complex medical image segmentation. By rigorously verifying the performance of the network in multimodal environments, this study not only provides an objective assessment of the practical application value of CMUeXt, but also provides critical theoretical support and practical guidance for future lightweight network optimization targeting CT/MRI data.

The structure of this article is as follows: Chapter 3 introduces the specific structure and implementation details of the CMUeXt network; Chapter 4 elaborates on experimental design and result analysis; Chapter 5 summarizes research conclusions and outlines future work. This research path aims to provide new insights and methods for model evaluation and optimization in the field of multimodal medical image segmentation.

2. Related work

The development of medical image segmentation technology has long centered on balancing accuracy and efficiency, especially for multimodal data such as CT and MRI. Early CNN-based models (e.g., 3D U-Net[4] and V-Net[5]) addressed volumetric data processing but incurred high computational costs, limiting real-time clinical applicability. Transformer-based architectures (e.g., TransBTS[6] and UNETR[7]) later improved long-range dependency modeling in multimodal fusion, yet their quadratic complexity and high annotation hampered practical deployment.

To optimize efficiency, lightweight networks have integrated dynamic convolution[8] and attention mechanisms[9], enhancing scalability and resource allocation for tasks such as organ and tumor segmentation . However, modal specificity (e.g., distribution shifts between CT and MRI) remains a key challenge. Domain adaptation and loss function adjustments (e.g., Focal Loss) have been proposed to address class imbalance and modal variability, though generalization gaps persist in cross-modal scenarios.

3. CMUNeXt Network Structure Reproduction and Modal Adaptation

3.1 Overall Network Architecture

CMUeXt is an efficient fully convolutional encoder-decoder network, consisting of five encoder layers and five decoder layers.. The encoder is responsible for downsampling and multi-scale feature extraction, while the decoder performs upsampling to gradually restore spatial resolution and ultimately output a segmentation map.

Each encoder layer comprises a CMUeXt block, a regular convolution block, and a downsampling operation. Unlike ResNet[10] or ConvNeXt, CMUeXt employs a regular convolution block consisting of 3x3 convolutions with a stride of 1, batch normalization (BN), and GeLU activation function at the top layer (Stem) of the network, which avoids resolution loss from initial downsampling and ensures consistency with top-layer skip connections.

Each decoder layer contains a Skip Fusion block and an upsampling block. The upsampling block consists of an upsampling layer, a convolutional layer, a batch normalization layer, and a GeLU activation function. CMUeXt employs bilinear interpolation to upsample feature maps by a factor of two. The convolutional layer adopts a 3 × 3 kernel with stride 1 and padding 1.

4. Experimental Design and Cross Modal Performance Analysis

4.1 Dataset and Preprocessing

(1) CT dataset:

The CT segmentation model was trained using the public Lung CT Nodule/Lesion Segmentation Dataset[11], which aggregates lung CT images for nodule segmentation. The dataset includes images from public repositories related to COVID-19, lung cancer, and other lung abnormalities. It contains 5069 images (including 2535 mask images and 2534 CT images) with a resolution of 512 × 512 pixels.

The CT model's generalization was verified using the public CT Liver Dataset [12], constructed for comparative liver segmentation research. Liver segmentation and morphological attribute (length, width, area) analysis support diagnosis, screening, treatment, and evaluation of liver diseases (e.g., hepatitis, liver tumors, autoimmune liver disease). The dataset includes 232 images (116 mask images and 116 CT images) with a resolution of 512×512 pixels.

(2) MRI dataset:

The MRI segmentation model was trained using the public BRAIN TUMOR MRI DATASET [13], which includes two subsets: a classification subset (glioma, meningioma, pituitary tumor, tumor-free) and a segmentation subset (tumor presence/absence). Only the segmentation subset was used in this study, containing 4384 images (2192 mask images and 2192 MRI images) with a resolution of 512×512 pixels.

The MRI model's generalization was verified using the public Brain Tumor Segmentation Dataset [14], which classifies brain tumors into four categories (glioma, meningioma, pituitary tumor, tumor-free). Only tumor-containing samples were selected, totaling 264 images (132 mask images and 132 MRI images, with one-to-one correspondence) and a resolution of 512×512 pixels.

(3) MIX dataset:

A mixed dataset was constructed by combining the training subsets of the CT and MRI datasets, and a corresponding model was trained on this mixed data.

All datasets used in the experiment include images with matching mask labels. Original images and mask images were sorted into separate folders, with filenames ensuring one-to-one correspondence. For training dataset, data was split into 70% for training and 30% for real-time validation to iteratively optimize the model. For generalization validation datasets, the pre-trained model was directly imported, and validation was performed via an independent program.

4.2 Experimental setup and evaluation criteria

Experiment was conducted on an NVIDIA RTX 4090 GPU with CUDA 11.8. The total training epochs were set to 250, with training parameters consistent with the original CMUeXt network. Intersection over Union (IoU) and F1-score were adopted as core segmentation performance evaluation metrics. For generalization validation, the pre-trained model was applied to new datasets, and segmentation results were visualized to assess performance.

4.3 Verification Results and Analysis

This section quantitatively evaluates CMUNeXt performance in CT and MRI segmentation tasks through multi-metric comparison, and visually identifies its shortcomings in complex medical image segmentation through new datasets.

Optimal segmentation models for CMUeXt and its two variants were trained on the CT, MRI, and MIX datasets, respectively. The model's train loss (train-loss), intersection to union ratio (val_IoU), validation F1 score (val_f1), and validation accuracy (val-ACC) were recorde., Visual validation was performed on the validation set using the optimal models obtained. Quantitative results are presented in Table 1:

Table 1. Result Comparison Table

	Train_Loss	Val_IOU	Val_F1	Val_ACC
CT CMUNeXt-S	0.1820	0.6648	0.7742	0.9937
CT CMUNeXt	0.1445	0.6888	0.7934	0.9941
CT CMUNeXt-L	0.1498	0.6889	0.7941	0.9942
MRI CMUNeXt-S	0.4995	0.6807	0.7757	0.9933
MRI CMUNeXt	0.4652	0.7101	0.8024	0.9939
MRI CMUNeXt-L	0.4633	0.7154	0.8068	0.9939
MIX CMUNeXt-S	0.3352	0.6596	0.7646	0.9943
MIX CMUNeXt	0.3052	0.6842	0.7851	0.9947
MIX CMUNeXt-L	0.2990	0.6975	0.7973	0.9948

Visual validation results of the segmentation models are shown in Figures 1, 2, and 3. These results indicate that the segmentation performance of the trained models is suboptimal, with specific limitations as follows:

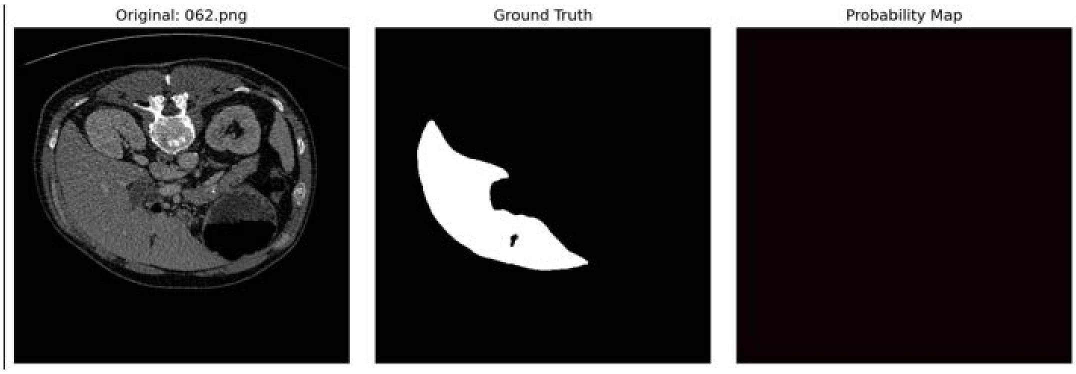

Figure 1. CT verification of CT model

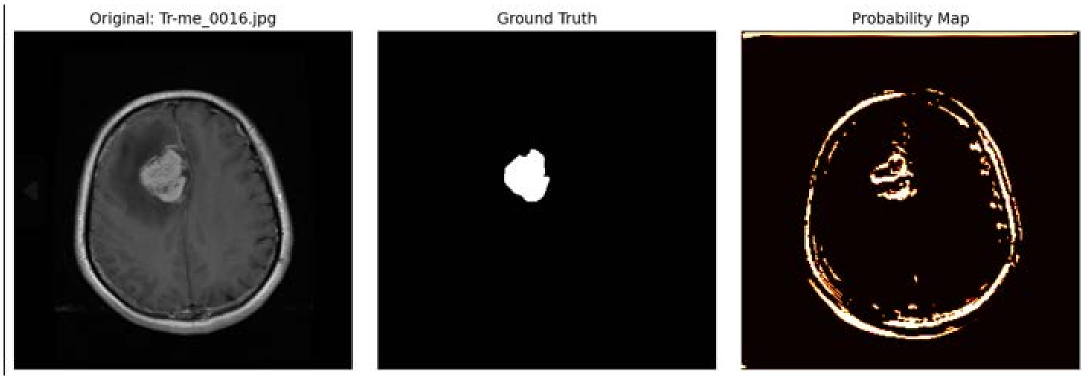

Figure 2. MRI verification of MRI model

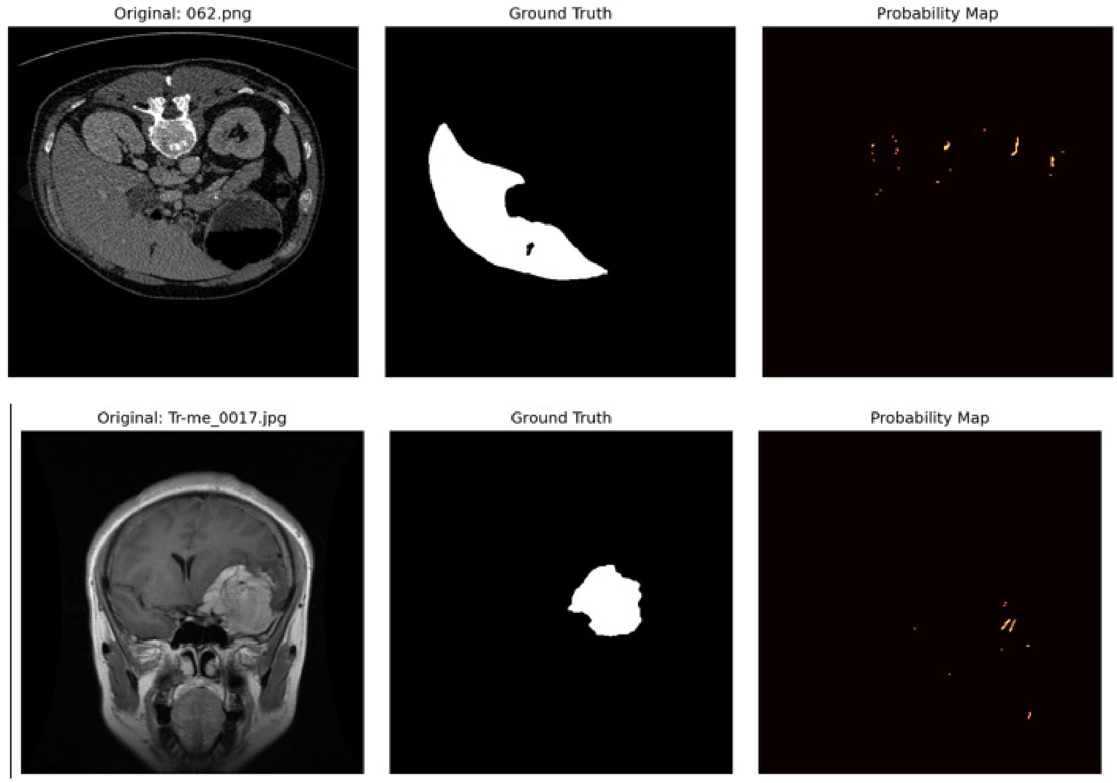

Figure 3. CT and MRI validation of MIX model

This experiment quantitatively evaluated and visually validated the segmentation performance of the CMUeXt model and its variants on CT, MRI, and mixed datasets (MIX). Analysis indicates that while quantitative metrics suggest the model achieved high segmentation accuracy, a significant discrepancy exists between visualization results and numerical indicators. This discrepancy implies that the observed high-precision values may stem from inherent biases in validation set construction or the evaluation process itself. Specifically, extremely high accuracy may result from class imbalance, as background pixels dominate medical images. Higher IoU and F1-score values may reflect potential data overlap between training and validation sets, enabling the model to achieve high scores by memorizing specific data rather than learning generalized features, which leads to inadequate generalization ability.

Further analysis of results across different imaging modalities revealed that the model exhibited lower training loss and relatively stable performance on CT data, while significantly higher training loss was observed on MRI and mixed datasets, which feature more complex textures and variable contrast mechanisms. This indicates that the model struggles to learn robust, discriminative feature representations from such data. Even on CT data, probability maps show weak response intensity in abdominal CT images. The model fails to accurately depict organ boundaries, resulting in blurred morphological features. For MRI brain tumor segmentation, the probability map exhibits a high degree of uncertainty, with inconsistent segmentation outputs between confidence thresholds, reflecting the model's inability to stably predict complex anatomical regions.. This visual evidence aligns with the quantitative metrics in Table 1, emphasizing the suboptimal performance of the CMUeXt architecture in processing multimodal data. While CMUeXt-L outperforms other variants slightly in most metrics, its overall improvement is limited and fails to address the model's fundamental segmentation flaws in complex scenarios. Overall, the results of this reverse validation demonstrate that the current model's actual segmentation performance and reliability require substantial improvement when facing complex medical images. This analysis underscores the need for architecture improvements, such as dynamic convolution mechanisms and attention-based fusion, to enhance cross-modal generalization in lightweight medical imaging networks.

5. Conclusion

This article aims to reproduce and validate the performance of the CMUeXt network in CT and MRI multimodal medical image segmentation tasks, revealing significant limitations of this lightweight architecture in complex medical scenarios. The experimental results show that, in CT chest and abdominal image segmentation, the network exhibits insufficient basic feature extraction ability, weak probability map response intensity, and difficulty in accurately capturing organ morphological features.; In MRI brain tumor segmentation, probability map analysisindicates high uncertainty in the model's judgment of tissue boundaries. Segmentation results lack stability across different confidence thresholds, and effective segmentation cannot be achieved in some cases. In-depth analysis identifies that the core design of CMUeXt, a large convolutional kernel structure, reduces boundary perception sensitivity in MRI's complex texture environments, and lacks systematic feature representation ability in CT images. The skip fusion module fails to effectively adapt to distribution differences in multimodal data, highlighting the insufficient generalization ability of lightweight networks in medical image applications. Based on these findings, future research should focus on architecture optimization, such as introducing dynamic convolution mechanisms to adaptively adjust receptive fields, integrating attention modules to enhance focus on key regions, and combining domain adaptation techniques to reduce feature distribution differences between CT and MRI; Additionally, the improved model should be validated on larger-scale, multi-center datasets, and the real-time requirements of edge deployment should be explored to promote the practical application of lightweight networks in clinical auxiliary diagnosis.

References

[1] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing.

[2] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. ACM computing surveys (CSUR), 54(10s), 1-41.

[3] Tang, F., Ding, J., Quan, Q., Wang, L., Ning, C., & Zhou, S. K. (2024, May). Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.

[4] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing.

[5] Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October).

V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV) (pp. 565-571). Ieee.

[6] Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021, September). Transbts: Multimodal brain tumor segmentation using transformer. In International conference on medical image computing and computer-assisted intervention (pp. 109-119). Cham: Springer International Publishing.

[7] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 574-584).

[8] Yang, B., Bender, G., Le, Q. V., & Ngiam, J. (2019). Condconv: Conditionally parameterized convolutions for efficient inference. Advances in neural information processing systems, 32.

[9] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11534-11542).

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[11] Piyush Samant. Lung CT nodule/ Lesion Segmenbtation. https://www.kaggle.com/datasets/piyushsamant11/pidata-new-names/data

[12] zxcv2022. CT liver. https://www.kaggle.com/datasets/zxcv2022/digital-medical-images-for--download-resource/data

[13] Bilal Açıkgöz. BRAIN TUMOR MRI DATASET. https://www.kaggle.com/datasets/bilalakgz/brain-tumor-mri-dataset

[14] Atika Akter11, Sabbir Ahmed. Brain Tumor Segmentation Dataset. https://www.kaggle.com/datasets/atikaakter11/brain-tumor-segmentation-dataset