IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 January 2024, accepted 29 January 2024, date of publication 6 February 2024, date of current version 12 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3360889

RESEARCH ARTICLE

MLU-Net: A Multi-Level Lightweight U-Net for Medical Image Segmentation Integrating Frequency Representation and MLP-Based Methods

LIPING FENG[1], KEPENG WU[2], ZIYI PEI[3], TENGFEI WENG[4], QI HAN[2], (Member, IEEE), LUN MENG[2], XIN QIAN[2], HONGXIANG XU[2], ZICHENG QIU[2], ZHONG LI[2], YUAN TIAN[2], GUANZHONG LIANG[5], AND YAOJUN HAO[1]

[1]Department of Computer Science, Xinzhou Normal University, Xinzhou, Shanxi 034000, China
[2]College of Intelligent Technology and Engineering, Chongqing University of Science and Technology, Chongqing 401331, China
[3]College of Materials Science and Engineering, Chongqing University of Arts and Sciences, Chongqing 402160, China
[4]School of Electrical Engineering, Chongqing University of Science and Technology, Chongqing 401331, China
[5]Chongqing University Cancer Hospital, Chongqing 400030, China

Corresponding author: Kepeng Wu (iwkp@cqust.edu.cn)

This work was supported in part by the West Light Foundation of the Chinese Academy of Science; in part by the Research Foundation of the Natural Foundation of Chongqing City under Grant cstc2021jcyj-msxmX0146 and Grant cstc2021jcyj-msxmX1212; in part by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant KJQN202301517, Grant HZ2021015, Grant KJZD-K202100104, and Grant KJQN202301543; in part by the Chongqing Science and Technology Military-Civilian Integration Innovation Project (2022); in part by the Bingtuan Science and Technology Program in China under Grant 2021AB026; in part by the Shanxi Province Applied Basic Research Program, China, under Grant 202203021211116; and in part by the Oil and Gas Production Safety and Risk Control Key Laboratory of Chongqing Open Fund under Grant cqsrc202110.

ABSTRACT Medical image segmentation is a challenging and popular task in the field of medical image processing in recent decades. Most of the current mainstream segmentation networks are based on convolutional neural networks (CNNs) methods. Among them, encoding and decoding structures based on U-Net architecture and skip connection mechanism have made great progress in medical segmentation. However, these networks come with increased complexity and training difficulty as the accuracy of network segmentation continues to increase, and their ability remains to be improved for extracting feature information in specific information-intensive structure segmentation tasks, such as brain tumors. In addition, the high training cost raises the application threshold of medical image segmentation. To address these issues, we introduce a frequency representation approach that can effectively reduce the loss of feature during encoding and decoding of segmentation networks. Then a tokenized multi-layer perceptron (MLP) method is introduced to learn the space information. Frequency representation and tokenized MLP can greatly reduce the parameters and computational effort while achieving more accurate and efficient medical image segmentation. Therefore, a multi-level lightweight U-Net segmentation network named MLU-Net is proposed to perform segmentation tasks of medical images quickly. In brain tumor segmentation experiments under equivalent preprocessing conditions, our network achieves substantial efficiency gains with parameter and computational workload reductions to 1/39 and 1/61 of U-Net's, while simultaneously demonstrating superior performance, enhancing the Dice and Intersection over Union (IoU) metrics by 3.37% and 3.30%, respectively. In addition, we perform experiments on dermatologic data and still achieve segmentation performance that outperforms comparable networks. These experiments show that the proposed network is characterized by lightweight and high accuracy, which is contributing to the exploration of clinical medicine scenarios.

INDEX TERMS Medical image segmentation, convolution neural networks, frequency representation, MLP-based, brain tumor segmentation, computer-aided diagnosis.

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

I. INTRODUCTION

Automatic segmentation of medical images holds paramount significance within the realm of medical imaging. Medical images encompass intricate structures and organizational information derived from diverse scanning modalities such as X-ray [1], CT scans [2], [3], and MRI [4], among others. Nevertheless, reliance on medical experts for diagnosis entails temporal and resource expenditures and is susceptible to clinical experience bias. Automated segmentation techniques precisely extract regions of interest (such as organs, anomalies, vessels) from medical images, effectively segregating them from backgrounds or other structures. This augmentation facilitates clinicians in diagnosis, treatment, and disease monitoring. Automated medical image segmentation facilitates easier discernment and analysis of characteristics such as shape, size, and spatial orientation of anomalies and tumors than manual observation. The meticulous extraction of regions of interest from intricate backgrounds or other structures significantly contributes to bolstering the accuracy and dependability of diagnostics and treatments [5]. As a computer-aided diagnostic system, automated medical image segmentation serves as an exceptional tool for aiding clinicians in gaining a deeper understanding of a patient's medical condition throughout the course of treatment.

Advancements in human proficiency and evolving medical exigencies impose heightened requisites upon medical practitioners. Distinct medical images encompass varying degrees of intricacy. As exemplified in Figure 1, Brain tumors exhibit an array of intricate configurations. What's clear is that human progress in medicine needs to be supported by improved computer-aided diagnostic systems.

Early methodologies for medical image segmentation predominantly relied on threshold-based techniques [6]. These approaches entailed the division of images into target and background through the selection of appropriate pixel intensity thresholds. While straightforward to implement, these methods exhibited suboptimal performance for images characterized by intricate backgrounds and noise. Subsequently, techniques emerged encompassing region growing, region splitting and merging, edge detection, contour evolution, and graph-theory-based segmentation. With the evolution of machine learning techniques, machine-learning-based approaches for medical image segmentation came to the fore, and emarkable advancements have been witnessed in the domain of medical image segmentation owing to deep learning techniques [7], [8]. Notably, the advent of convolutional neural networks(CNNs) in recent years has substantially augmented the semantic segmentation capabilities of medical images. Deep learning methods have harnessed data-driven training to apprehend the inherent features and contextual information within medical images, consequently yielding precise segmentation outcomes.

However, it is noteworthy that these computational procedures are primarily operated within the spatial domain, without due consideration of the disparities between computer vision and human visual perception. With the objective of extracting enhanced information from images, we introduce image processing methodologies rooted in the frequency domain representation, which can be processed by a computer. The amalgamation of spectral maps and spatial diagrams is employed to jointly extract characteristic information, thereby effecting an amelioration in the performance of medical image segmentation.

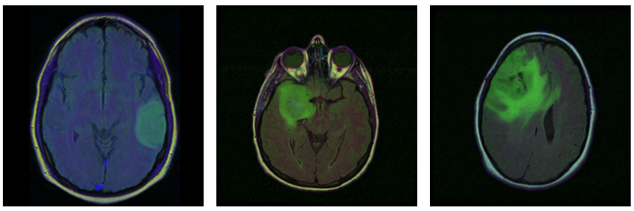


FIGURE 1. Some samples of brain tumors, where the irregular green parts are regions of tumor lesions in the brain.

Mainstream methods in medical image segmentation networks commonly adopt an encoder-decoder architecture. Seg-Net [9], acknowledged as the pioneer in the encoder-decoder segmentation network paradigm, remains widely embraced. Meanwhile, the predominant landscape of medical image segmentation networks predominantly revolves around the U-Net architecture and its variants [10]. Skip connections introduced innovatively in U-Net's between the encoder and decoder facilitates enhanced fusion of high-level and low-level features, thereby significantly augmenting the capacity for learning distinctive features pertinent to segmentation targets [11]. This advancement has notably propelled groundbreaking strides within the realm of medical image segmentation. In contrast to alternative architectures, U-Net manifests expedited processing and superior efficacy in segmenting medical images, consequently spawning a proliferation of U-Net-based enhancements in recent years, including U-Net++ [12], [13] and related derivations. U-Net has indeed emerged as the foundational bedrock underpinning nearly all mainstream methodologies in medical image segmentation [14]. However, as classic convolutional neural network models, U-Net and its variant networks continue to grapple with several inherent challenges:

1) The utilization of small convolution kernels throughout the entire convolutional process imposes a local context constraint on the network's capacity for feature acquisition. Furthermore, the extensive convolutional kernel operations throughout the network engender substantial computational complexity and parameter proliferation.

2) During the downsampling phase, the predominant adoption of convolution and pooling techniques in U-Net processes both critical and non-critical information uniformly within the spatial domain, resulting in limited preservation of valuable feature information.

3) The U-Net architecture exhibits a simplistic superposition of information during upsampling and the decoding process with skip connections, failing to adequately amplify

the information content of high-resolution but low-level semantic feature maps and low-resolution but high-level semantic feature maps.

In response to this array of challenges, we have introduced frequency domain-based upsampling and downsampling techniques into the foundational U-Net structure. Furthermore, we have enhanced the network architecture through the integration of MLP methodologies. These strategic enhancements optimize the network's capacity for global feature information acquisition while concurrently mitigating network complexity and computational demands.

Finally, we construct an efficient and lightweight network architecture. In summary, the principal contributions of our study can be encapsulated as follows:

- A refined frequency representation approach is introduced into CNNs and is specifically employed in the context of downsampling and upsampling procedures, demonstrating heightened efficiency in acquiring enriched semantic information relevant to segmentation objectives. Modules built upon this methodology can seamlessly replace extant deep learning modules without necessitating significant adjustments.
- In pursuit of network lightweighting and the acquisition of more enriched feature information, the conventional U-Net architecture is supplemented with a tokenized MLP. Therefore, a lightweight network, MLU-Net, is proposed, which integrates both frequency domain information and tokenized MLPs.
- MLU-Net attains superior segmentation effect in contrast to existing state-of-the-art networks, while employing a mere 1/39 of the parameter count and a 1/61 of computational load comparing that to U-Net, all within identical preprocessing conditions. Furthermore, MLU-Net demonstrates its adaptability across diverse medical segmentation scenarios, wherein it exhibits powerful performance and robustness.

II. RELATED WORK

In this section, relevant work of medical image processing will be reviewed. Meanwhile, methods of frequency domain representation, medical image segmentation networks and MLP-based methods are elaborated and analyzed.

A. PROCESSING OF FREQUENCY DOMAIN IMAGES

Frequency domain images refer to images represented in the frequency domain through Fourier transformation or other frequency domain transformation methods, which transition images from the spatial domain to the frequency domain. In the frequency domain, high-frequency components signify image details and variations, while low-frequency components denote overall image structures and approximate shapes. Frequency domain images provide an alternative perspective for comprehending and processing images. Analyzing an image's frequency distribution facilitates the acquisition of richer information, thereby yielding enhanced outcomes in image processing and analysis.

In order to translate spatial domain images to frequency domain images, two-dimensional discrete Fourier transform finds extensive application in the realm of image processing. Fourier transform furnishes a potent representation for feature extraction [15], offering a convenient means to acquire various types of feature information. Leveraging this characteristic, the 2D discrete Fourier transform can be more effectively integrated within convolutional neural networks. Numerous scholars have embarked on innovative explorations within convolutional neural networks. Pratt et al. [16] introduced the Fourier Convolutional Neural Network (FCNN), yielding satisfactory performance. Ayat et al. [17] proposed the spectral rectified linear unit (SReLU) as an activation function to address issues posed by computationally intensive domain transformations. Pour and Seker [18] advocated integrating transform domain representation by injecting Laplacian pyramids into the network architecture, its performance in image processing substantiates its promising utility [15], [19].

In this paper, our intention is to segregate low-frequency contour information from high-frequency detail information within the frequency domain, due to information-rich characteristics and information distribution attributes of spectrogram. Subsequently, we plan to integrate this segregated information into the upsampling and downsampling processes of medical image segmentation networks.

B. NETWORK MODEL OF MEDICAL IMAGE SEGMENTATION

In the field of medical image segmentation, U-Net is a widely recognized CNN model proposed by Ronneberger et al. [11], where this network contains encode-decode structure and skip connections. The encoding part employs a contracted path to capture contextual information, while the decoding part uses a symmetric expanded path to achieve precise localization of the segmentation targets. The network is designed to process the entire image end-to-end, directly generating the segmentation map. Owing to its efficacious skip connection mechanism, which integrates high-level low-resolution features with low-level yet high-resolution features, the U-Net architecture has garnered extensive adoption in recent scholarly endeavors [12], [20], [21], [22], [23], [24], [25].

In an extended inquiry into the skip connection mechanism, Zhou et al. [12] and Zhang et al. [22] innovatively devised novel skip paths with the intent of diminishing the semantic and resolution disparity that exists between low-level and high-level features. Numerous complementary approaches have been introduced in this realm, including the R2U-Net method developed by Alom et al. [20], the DENSE-INception U-Net framework pioneered by Zhang et al. [26], the attention U-Net architecture advanced by Oktay et al. [27], and the bi-directional ConvLSTM U-Net

model proposed by Azad et al. [21]. Notably, Milletari et al. introduced V-Net [25], which represents a three-dimensional extension of the U-Net architecture, enabling the segmentation of volumetric data in a single pass. It is imperative to acknowledge that as these networks contribute to the refinement of segmentation accuracy, they also notably engender escalated computational demands and intricacy within the network architecture. Within the confines of this study, we introduce techniques centered around frequency representation and MLP methodologies into the framework of the U-Net architecture. This amalgamation culminates multi-level convolutional neural network tailored for medical image segmentation, denoted as the multi-level U-Net (MLU-Net).

C. MLP-BASED METHODS

The resurgence of interest in the realm of computer vision is currently being catalyzed by the emergence of the MLP, a multi-layered neural architecture shown in (a) of Figure 2. This resurgence has been notably accentuated by the recent proposition of the MLP-Mixer by the Google research collective [28]. This architectural innovation, characterized by its token-mixing and channel-mixing modules, represents a departure from conventional convolutional neural networks (CNNs), heralding a fresh avenue of exploration in feature extraction methodologies. As such, the discourse surrounding the MLP-Mixer underscores its potential to augment the prevailing paradigms of image classification and analysis, inviting scholarly attention and conjecture regarding its implications for advancing the frontier of computer vision.

MLP-Mixer was proposed by Tolstikhin et al. [28] in 2021. The architecture is based on the idea of using only MLPs for all computation, without using any convolutional layers. It consists of two types of layers, called the ''channel-mixing layer'' and the ''token-mixing layer''. The channel-mixing layer operates on each channel of the input tensor independently, while the token-mixing layer operates on each spatial location (or ''token'') of the tensor independently. By using only MLPs, the MLP-Mixer architecture is more flexible and scalable than traditional CNNs. It also achieves state-of-the-art performance on several image classification benchmarks while requiring fewer computational resources. Overall, MLP-Mixer represents a promising direction for the design of neural network architectures for computer vision. The method achieves similar segmentation results as ViT [29] on some mainstream datasets.

However, the MLP-Mixer framework has limited generalization capability and often misses low-level semantic information. To address these limitations, the AS-MLP proposed by Lian et al. [30] is designed to address the issue of shift-variance in traditional methods. It consists of a stack of axial shifted multi-layer perceptrons, which are MLPs with shift operations applied in the axial direction. Specifically, the input feature maps are divided into multiple groups, and each group is shifted along a specific axis independently. The architecture also includes a ''coarse-to-fine'' strategy, where the lower layers of the network capture coarse features, while the higher layers capture finer features. Its structure is shown in (b) of Figure 2. This strategy is achieved using multiple MLPs with different kernel sizes and feature map sizes.

AS-MLP achieves state-of-the-art performance on several image classification benchmarks, and also outperforms other shift-equivariant networks on small image datasets. It is computationally efficient and can be easily implemented in existing deep learning frameworks, making it a promising direction for the design of neural network architectures for computer vision.

Valanarasu and Patel [31] have developed a novel framework called UNeXt by incorporating tokenized MLP blocks into U-Net. This framework is designed to extract local-to-global semantic information from input feature maps by performing sequential feature transformations on the vertical and horizontal dimensions, corresponding to different axial shifts. Unlike AS-MLP, which simply sums features in two dimensions, UNeXt employs a symmetric encoder-decoder architecture and can effectively reduce model parameters and time complexity. The structure of tokenized MLP is shown in (c) of Figure 2. Here, tokenized MLP will be introduced into the fourth and fifth level of our proposed MLU-Net architecture to focus on high-level semantic information.

III. METHOD

In this section, an overview of our proposed network is first given. Then, the details of each level of the proposed multi-level network will be explained.

A. MLU-NET

The whole network architecture of MLU-Net is shown in the Figure 3, which is a multi-level segmentation network based on U-Net. Given an input image $I \in \mathbb{R}^{H\times W\times C}$, where H, W and C are the height, width, and channel number of the input images in the network. The channel number of feature map in each layer of U-Net is decreased, in order to reduce the number of network parameters and the computational complexity, and solve the overfitting problem in training. Then we propose substituting the original channel numbers (C1, C2, C3, C4, C5) in the U-Net architecture with lower values. This modification significantly enhances the network's training efficiency while preserving excellent metrics in medical image segmentation.

The MLU-Net architecture consists of five hierarchical levels, exhibiting distinct processing strategies, in addition to varying scales and channel numbers. The first three levels employ a CNN approach with reduced channel numbers, where each level of the encoding and decoding structure has 2D convolutional(Conv2D) layers, Rectified Linear Unit (ReLU) layers, and batch normalization(BatchNorm) layers. The encoding structure in these three levels utilizes a multi-downsampling module to reduce the dimensions of feature maps, which is explained in detail in section III-C and replaces traditional pooling methods while preserving more comprehensive essential feature information. Similarly, in the

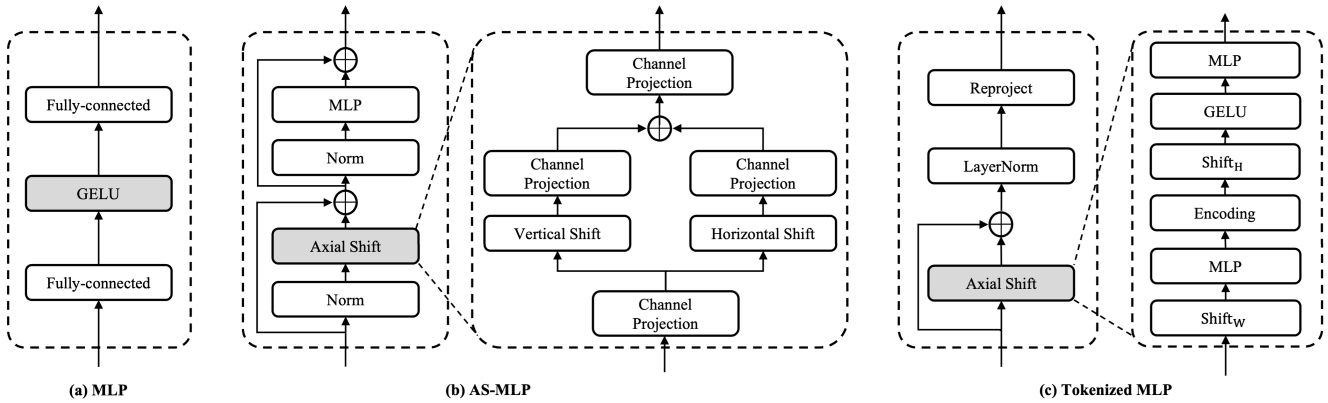


FIGURE 2. MLP and its evolved MLP-based axial shift structures.

decoding structure, a frequency domain upsampling module is incorporated before each CNN level. The spectrogram upsampling block(SUB) replaces conventional interpolation upsampling methods to restore finer feature information, as elaborated in section III-D. Additionally, skip connections between the same hierarchical levels are accomplished through the collaborative efforts of the multi-downsampling module and the spectrogram upsampling block. In the fourth and fifth levels among the five hierarchical levels, an MLP-based approach is employed. This results in a novel multi-level network composed of various types of layers, which we term as MLU-Net.

B. APPLICATION OF FOURIER TRANSFORM

Presently, the majority of research in the realm of medical image processing tends to focus on spatial domain methodologies. Spatial domain imagery aligns with the visual data captured by the human visual system, thereby rendering the information more readily perceptible and comprehensible to human cognition. However, disparities exist between the cognitive processes of computer vision and the human visual system in the interpretation of images. Humans frequently rely on contextual and situational information to imbue images with deeper meanings during their comprehension. In contrast, computer vision systems often encounter limitations in handling context and situational nuances due to the fixed nature of input data, thereby hampering the comprehensive consideration of dynamic factors in task design [32]. In the design of computer vision tasks, it is crucial to acknowledge that humans tend to dynamically adjust sensitivity to context when processing images, a nuance that may be lost in static computer vision tasks. Consequently, designing computer vision tasks based on our own cognitive paradigms inevitably results in the loss of cognitive information. Employing frequency domain processing methods can mitigate this information loss to some extent.

Considering the above, we further introduce the concept of frequency domain representation of images. Transforming spatial domain images into spectrums of the frequency domain, and then the spectrum graph is more suitable for distinguishing between structural and detailed information in an image. As opposed to conventional spatial domain methods, frequency domain entails a more abstract depiction of image information. Each pixel within the frequency domain spectrum corresponds to distinct frequency band image details, and their collective fusion constitutes the entirety of the spectrum's information. In the spatial domain, individual pixels merely convey localized details, and it is the aggregation of these local details that forms a comprehensive spatial image. Given that medical image acquisition predominantly occurs via signal-based instrumentation, the advantages of processing medical images in the frequency domain are evident. To sum up, the frequency domain representation approach presents substantial opportunities for advancement of medical image processing.

The process of converting spatial domain images into frequency domain representations is notably achieved through the utilization of the two-dimensional discrete Fourier transform(2D-DFT). The versatility of the Discrete Fourier Transform (DFT) extends to applications in both two-dimensional and higher-dimensional contexts. Given a spatial domain image denoted as $f(x, y) \in \mathbb{R}^{H\times W}$, the two-dimensional discrete Fourier transform yields the frequency domain representation in the form of the spectrum $F(u, v) \in \mathbb{R}^{H\times W}$. This transformation and inverse transformation can be formally expressed as:

$$F(u, v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} f(x, y)\, e^{-2\pi i(\frac{xu}{H}+\frac{yv}{W})} \tag{1}$$

$$f(x, y) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} F(u, v)\, e^{2\pi i(\frac{ux}{H}+\frac{vy}{W})} \tag{2}$$

here, H and W symbolize the dimensions of height and width, while $u,\ x \in \{0, \ldots, H\text{-}1\}$ and $v,\ y \in \{0, \ldots, W\text{-}1\}$ respectively denote the spatial coordinates within the image domain. Its inverse transformation format is expressed as $F^{-1}(u, v) = F^*(u, v)$, representing the complex conjugate

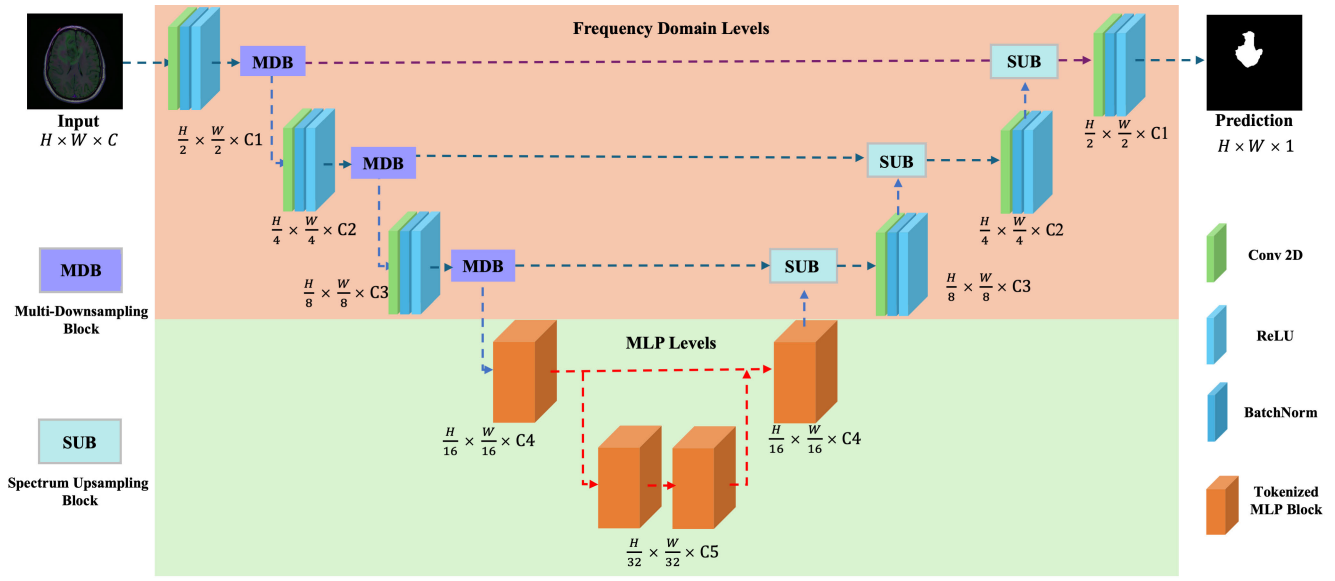


FIGURE 3. Overview of MLU-Net, a multi-layer lightweight U-Net structure network with frequency domain levels and MLP levels. It mainly contains multi-downsampling module(MDB), spectral upsampling module(SUB), tokenized MLP module to improve the network.

of the original transformation. A comprehensive elucidation of this concept is provided in article [15]. Leveraging this property, a distinctive approach to image processing, divergent from the conventional spatial domain, becomes attainable. Furthermore, a technique known as fast Fourier transform (FFT) exploits the distinguishing characteristics of the discrete Fourier transform-oddness, evenness, imaginary, and real-to refine the algorithmic aspects of the discrete Fourier transform. Ultimately, we opt to employ the FFT methodology to significantly reduce computational time complexities. The characteristics of Fourier transform and spectral image make image processing expand more abundant processing means, such as multiple images can be obtained for processing according to high and low frequency separation, and can be restored to the original image. Further, we can isolate the key information we need by different frequencies. The visualization of the Fourier transformation effects is illustrated in Figure 4.

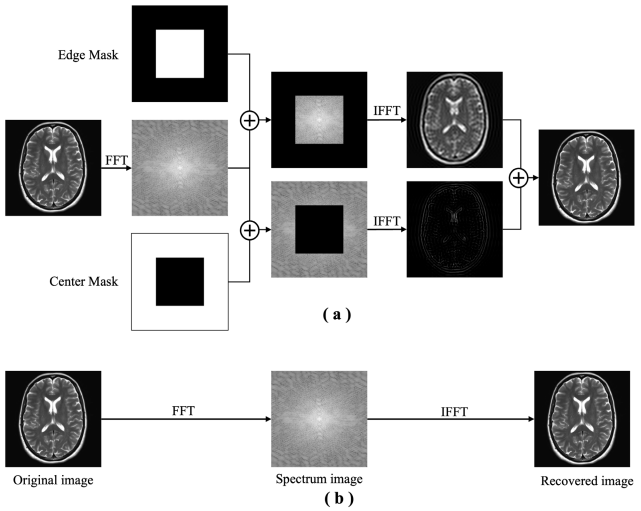


FIGURE 4. (a): Schematic diagram of high frequency and low frequency separation and reorganization effect. (b) The fast Fourier transform and the inverse fast Fourier transform are applied directly to the image. Through the fast Fourier transform of the image to obtain the spectral map, according to the high frequency and llow frequency a spectral map will be separated into two spectral maps by use of different mask, and then through the inverse fast Fourier transform to restore the original image. The effect is equivalent to the direct inverse fast Fourier transform.

C. MULTI-DOWNSAMPLING BLOCK(MDB)

Downsampling assumes a pivotal role within the convolutional neural network. It involves the reduction of spatial resolution in the input image while simultaneously augmenting the channel dimensions of feature maps. This synergy facilitates the assimilation of diverse-scale feature information from the input image, enhancing the network's capacity to comprehend both structural intricacies and contextual nuances within the image. Moreover, downsampling contributes to the mitigation of computational demands and memory footprint in image processing tasks. The progressive diminution of feature map dimensions engendered by this iterative process empowers the network to holistically perceive variations in the scale of input images. This attribute becomes particularly advantageous in addressing tasks like semantic segmentation, as it facilitates the more effective capture of multi-scale information. Presently, mainstream downsampling methods include techniques such as average pooling, max pooling, and probabilistic pooling. These downsampling methods selectively discard information from adjacent pixels in the spatial domain, employing similar strategies for both information-dense and sparse regions, thereby resulting in substantial information loss. The capacity

to retain valuable information during the downsampling process warrants further investigation.

The U-Net segmentation network employs a contraction path in its encoding structure, incorporating multiple downsampling operations to preserve pivotal feature information. In addition to convolutions, downsampling methods primarily entail the aggregation of local pixel values into representative one specific pixel value. These methods aim to concurrently diminish image resolution while retaining holistic information, commonly encompassing strategies such as max pooling and average pooling. However, these methods are rigid in their information loss as the resolution decreases. Addressing this limitation, we introduce a frequency domain downsampling approach, culminating in the proposal of a multi-downsampling block, which allows for selective preservation of feature details during the downsampling process. Notably, frequency domain spectrums exhibit distinct distributions of high and low-frequency image information. High-frequency components encompass intricate texture details, while low-frequency regions encapsulate essential contour information. The function of the multi-downsampling is to capitalize on this inherent characteristic [15]. In the context of medical image segmentation tasks, our focus centers on extracting the pivotal central low-frequency domain from the spectrum.

Subsequently, we delve into the comprehensive exposition of the novel multi-downsampling block. Multi-downsampling block comprises two distinct branches. The initial facet is rooted in the introduction of a downsampling approach grounded in the frequency domain representation, emphasizing the retention of the central low-frequency domain within the frequency spectrum to glean salient feature information. Commencing with the original input image of dimensions $H \times W$, fast Fourier transform yields a frequency spectrum of equivalent dimensions to those before the transform. Notably, a copy of this spectrum will be transmitted to the same layer in the decoding structure by way of a skip connection. A subsequent operation involves cropping the central low-frequency domain from the frequency spectrum. Specifically, the central region corresponding to one-fourth of the original image area is retained, resulting in a new frequency spectrum of dimensions $\frac{H}{2} \times \frac{W}{2}$. The application of a inverse fast Fourier transform(IFFT) facilitates the restoration of the new frequency spectrum to its original spatial domain dimensions of $\frac{H}{2} \times \frac{W}{2}$, yielding a downsampled low-resolution feature map. The algorithm of multi-downsamping is shown in Algorithm 1.

Within the Algorithm 1 workflow, the input is a feature map X of size $H \times W$. FFT denotes the fast Fourier transform, employed to generate a spectrum image of the same dimensions as the input image, as delineated in Equation (1). Furthermore, a duplicated sample, denoted as X_{Copy}, originating from X_{Spec}, is propagated as output and input into the corresponding upsampling block within the skip connection structure. The *Crop* operation entails the central region cropping of the input image, specifically,

Algorithm 1 Multi-Downsamping

Require: Feature map X of size $H \times W$
Ensure: Skip conection $X_{Copy} \in \mathbb{R}^{H \times W}$ and sampled image $\hat{X}$ of size $\frac{H}{2} \times \frac{W}{2}$
1: $X_{Spec} \in \mathbb{R}^{H \times W} \leftarrow FFT(X)$
2: $X_{Copy} \in \mathbb{R}^{H \times W} \leftarrow Copy(X_{Spec})$
3: $X_{Center} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}} \leftarrow Crop(X_{Spec})$
4: $X_{S1} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}} \leftarrow IFFT(X_{Center})$
5: $X_{S2} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}} \leftarrow Pooling(X)$
6: $\hat{X} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}} \leftarrow Add(X_{S1}, X_{S2})$
7: **return** $X_{Copy}, \hat{X}$.

a quarter of its size from the center, encompassing the height range from $\frac{1}{4}H$ to $\frac{3}{4}H$ and the width range from $\frac{1}{4}W$ to $\frac{3}{4}W$. Correspondingly, $IFFT$ represents the inverse fast Fourier transform operation as delineated in Equation (2). While *Pooling* signifies the utilization of the Max-Pooling method in this context. Finally, the *Add* operation combines the information from two image. Ultimately, the output is a spectrogram backup X_{Copy} of size $H \times W$ and a downsampled feature map $\hat{X}$ of size $\frac{H}{2} \times \frac{W}{2}$.

In comparison with conventional pooling methods, this frequency domain-integrated approach selectively discards feature information with lower relevance to the task, such as textural details. Furthermore, recognizing the contextual limitations inherent in this characteristic across diverse image types, we adopt a nuanced approach that amalgamates both frequency domain and conventional methods within a Multi-downsampling block. This amalgamation serves to augment the block's generalizability. Thus, emerges the Multi-downsampling block, as visually demonstrated in Figure 5.

D. SPECTRUM UPSAMPLING BLOCK(SUB)

In the context of medical image segmentation tasks, the role of upsampling is equally of paramount significance. The objective of medical image segmentation involves the discrimination of distinct structures or entities within an image, such as demarcating tumor regions in magnetic resonance images. However, the meticulous delineation of intricate details and contours within the image necessitates high-resolution feature maps. The process of upsampling is typically effectuated through methods such as interpolation or transposed convolutional layers, aiming to restore the lower-resolution feature maps to their original image dimensions, concurrently striving to reintegrate lost fine-grained details. This step constitutes a pivotal element within segmentation networks, wherein feature maps are elevated to the resolution of the original image.

In our approach, we contemplate a strategy for image detail recovery, whereby we introduce the adoption of an upsampling methodology rooted in the frequency domain representation [33]. This strategy selectively reconstitutes image details during the restoration of image details,

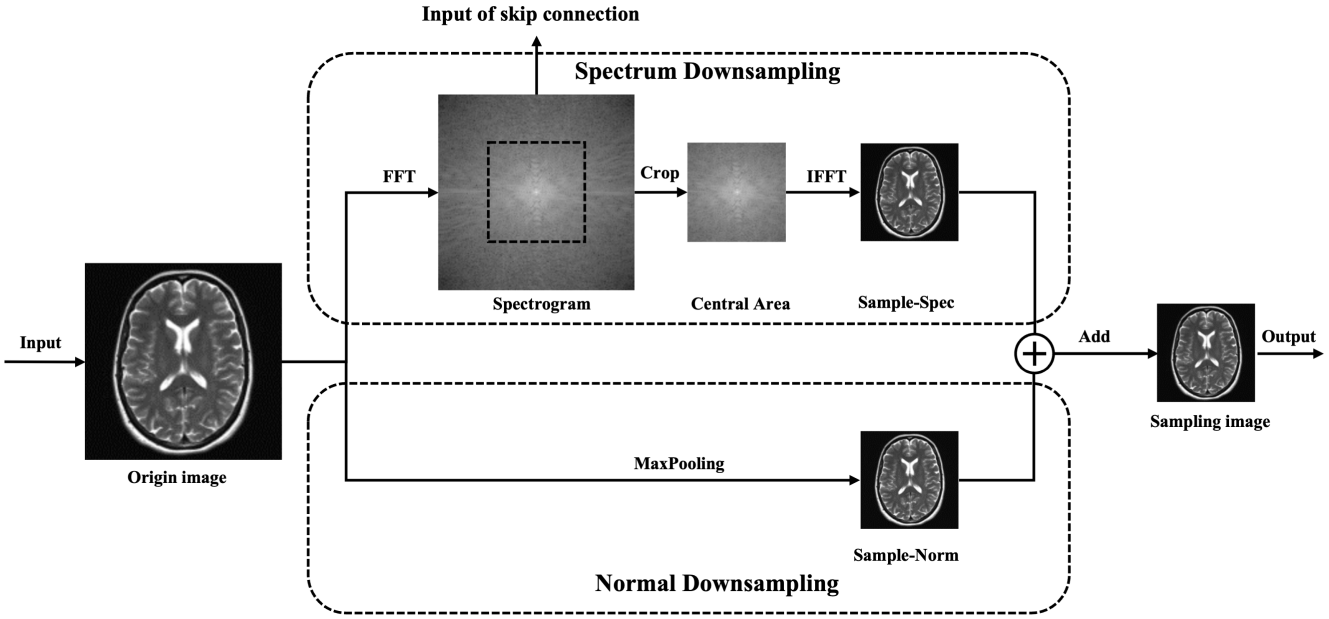


FIGURE 5. Multi-downsampling block(MDB), it consists of two parts: frequency domain downsampling and traditional downsampling.

embodying a distinctive attribute within the upsampling procedure.

The amalgamation of high-level semantic cues with low-resolution features and low-level semantic cues with high-resolution features constitutes an efficacious approach for the restoration of intricate target object details. The fusion of semantic information and high-resolution details in the frequency domain yields more potent connections than those achieved in the spatial domain. In this segment, we embrace an alternative strategy for upsampling, which involves the introduction of a method based on frequency domain representation [33], thereby substituting conventional techniques such as interpolation and transposed convolution.

To attain a high-resolution image encompassing fused high-level semantic features, a sequential procedure involves enhancing the dimensions of the spectral representation of high-level semantic features and subsequently embedding them within the high-resolution imagery of low-level semantic information. The algorithm of spectrum downsamping is shown in Algorithm 2.

The specific steps of Algorithm 2 are as follows: The high-resolution frequency domain spectrum feature map X_{Copy} with dimensions $2H \times 2W$, obtained from the skip connection, undergoes a Centre_Wipe operation by cropping to retain only the edge high-frequency region. This high-frequency portion of the spectrum map contains rich image detail information. This component is shown in the upper branch of Figure 6. Subsequently, the semantic high-level feature map X with dimensions $H \times W$ is transformed into a frequency spectrum feature map through fast Fourier inverse transform for further processing. This part includes abundant image contour information. The feature map's size

Algorithm 2 Spectrum Upsamping

Require: High-level semantic information feature map X of size $H \times W$, and high resolution spectrum feature map X_{Copy} of size $2H \times 2W$

Ensure: Upsampled feature map $\hat{X}$ of size $2H \times 2W$

1: $X_{Edge} \leftarrow Center\_Wipe(X_{Copy})$
2: $X_S \leftarrow FFT(X)$
3: $X_{Center} \leftarrow Edge\_Padding(X_S)$
4: $X_{Add} \leftarrow Add(X_{Edge}, X_{Center})$
5: $\hat{X} \leftarrow IFFT(X_{Add})$.
6: **return** $\hat{X}$.

is increased to $2H \times 2W$ by zero-padding along the edges, where the zero-padding process does not introduce additional information to the spectrum map. It corresponds to the lower half of the merge branch in Figure 6. Finally, the high-level semantic spectrum feature map is integrated with the high-resolution spectrum feature map from the skip connection to combine high and low frequency regions, resulting in a new upsampled frequency spectrum feature map with dimensions $2H \times 2W$. The spatially upsampled feature map $\hat{X}$ with dimensions $2H \times 2W$ is obtained through fast Fourier inverse transform. This completes the frequency domain upsampling process, as illustrated in Figure 6.

E. TOKENIZED MLP MODULE

In order to comprehensively capture spatial relationships among features while maintaining a lightweight network model, we introduced the tokenized MLP method [31] based on MLP in the fourth and fifth layers of MLU-Net.

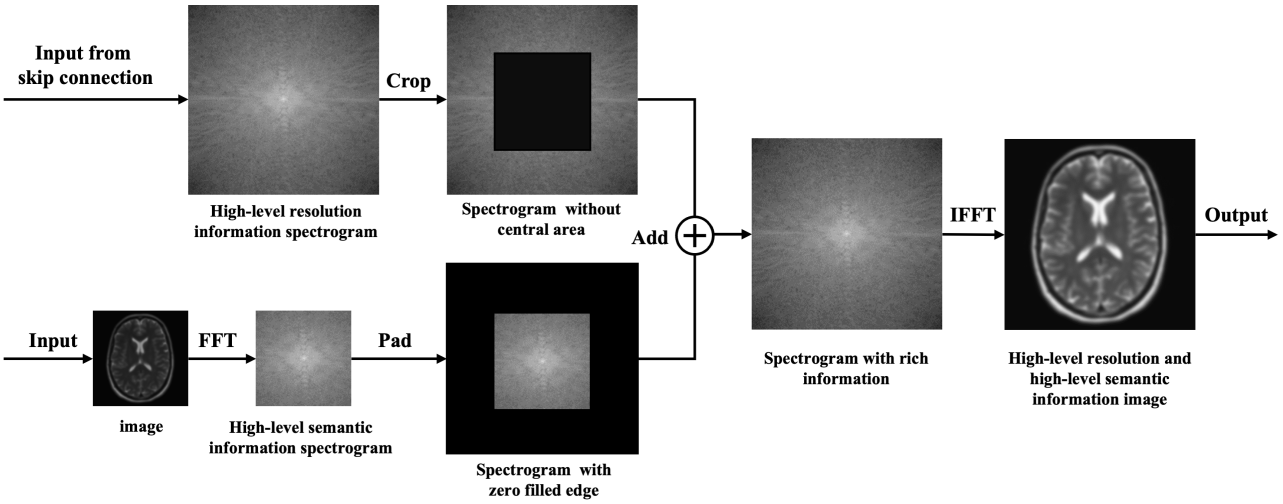


FIGURE 6. Spectrum upsampling block(SUB), fuses high-level semantic features and high-resolution features in spectral domain.

Traditional MLPs often demand a substantial number of parameters for processing image data within fully connected layers. However, the incorporation of axial shift mechanisms allows for more efficient information processing, alleviating the burden of excessive parameterization. The positional information introduced by axial shift operations facilitates a more profound understanding of the relative positions and spatial relationships among features. Consequently, the network is better at capturing features on different locations, thereby expanding its receptive field and enabling a more comprehensive comprehension of structural information within images. We integrated this approach into MLU-Net to enhance global information acquisition capabilities, thereby improving segmentation accuracy while simultaneously reducing the parameter count and computational complexity.

The tokenized MLP module takes the feature map X as input, and within the module, we employ an axial shift mechanism for spatial interaction on the input feature map X. The axial shift mechanism, the same as AS-MLP [30] (as shown in Figure 2(b)), is distinctive in tokenized MLP as it sequentially utilizes the shift mechanism in both horizontal and vertical directions. The feature X is partitioned into multiple distinct segments, and after shift the partitions horizontally and tokenize them, global feature information is extracted through MLP, utilizing depth-wise separable convolution (DWConv) for feature extraction. Similar procedures are applied in the vertical direction. This process effectively preserves salient feature information in the feature map while discarding features with low task relevance. Following the axial shift mechanism, there is a residual connection and layer normalization (LN) to enhance the model's generalization ability. The final output is reprojected to a dimensional consistency with the input X (as shown in Figure 2(c)). Reproject is employed to maintain dimensional consistency. The module of tokenized MLP is described by the following equations,

$$X_{\text{shift}} = \text{Shift}_{\text{W}}(X);\ T_W = \text{Tokenize}\left(X_{\text{shift}}\right), \tag{3}$$

$$Y = f\left(\text{DWConv}\left(\left(\text{MLP}\left(T_W\right)\right)\right)\right), \tag{4}$$

$$Y_{\text{shift}} = \text{Shift}_{\text{H}}(Y);\ T_H = \text{Tokenize}\left(Y_{\text{shift}}\right), \tag{5}$$

$$Y = f\left(\text{LN}\left(T + \text{MLP}\left(\text{GELU}\left(T_H\right)\right)\right)\right). \tag{6}$$

In equations (3), (4), (5) and (6), X represents the input feature map of the tokenized MLP module. Shift_{W} and Shift_{H} denote horizontal and vertical shift operations, respectively [30]. The X_{shift} and Y_{shift} are generated by Shift_{W} and Shift_{H}, respectively. T denotes the feature tokens of feature map X, and T_W and T_H represent of X_{shift} and X_{shift}, respectively. Correspondingly, the Reproject block in Figure 2(c), represented by f, is responsible for reprojecting the feature tokens to restore the feature map. In addition to the blocks represented in Figure 2(c), the Reproject block is also used before the second tokenization. Additionally, we incorporate the Gaussian error linear unit (GELU) for its smoothness, aiming to enhance the convergence speed and overall performance of the model [34]. Furthermore, layer normalization (LN) and residual connection are employed to ensure network stability. The final feature map Y is obtained by reprojection of the output feature tokens. The entire process is illustrated in Figure 2(c), emphasizing the transformation between feature maps and feature tokens as described in the depicted equations.

Considering that these computations occur in the embedding dimension E, which is notably smaller than the dimension of the feature map $\frac{H}{N} \times \frac{H}{N}$, where N is a factor determined by the module (typically 2), we adhere to the methodology inherited from paper [31]. Unless otherwise specified, E is set to 768. This tokenized MLP module design is instrumental in encoding valuable fea-

ture information, with the added benefit of not incurring additional burdens in terms of parameters and computational complexity.

F. LOSS FUNCTION

The choice of the Binary Cross-Entropy Loss with Logits as our loss function is rooted in its efficacy in estimating the similarity between predicted and actual segmented images during the training phase. This selection proves particularly advantageous when confronted with imbalanced datasets, a prevalent characteristic in medical imaging scenarios where instances of pathology constitute a minority class. Diverging from the conventional Binary Cross-Entropy Loss, this variant operates on unprocessed values (logits) generated by the model, as opposed to probability values, thereby offering stability and efficiency in optimization. The incorporation of logits mitigates the vanishing gradient problem, thereby enhancing convergence during training. In the context of medical image segmentation, this loss function exhibits notable advantages in terms of numerical stability. Binary Cross-Entropy Loss with Logits is represented by the equations:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\sigma(\hat{y}_i)) + (1 - y_i) \cdot \log(1 - \sigma(\hat{y}_i))\right] \tag{7}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

where $\hat{y}$ denotes the raw output predicted by the model, while y represents the actual binary labels. The symbol σ signifies the sigmoid activation function shown in Equation (8), and N corresponds to the batch size, which signifies the number of samples processed in a single training iteration. This formula signifies the computation of the average loss across all samples within the batch, where the ultimate loss value, denoted as L represents the mean loss per sample. This metric serves as a performance indicator for the model and guides the backpropagation process to update model parameters. A higher similarity between predicted results and ground truth leads to lower loss function values, indicative of superior predictive performance.

IV. EXPERIMENTS

In this section, the performance of the model MLU-Net is evaluated in experiments. We compare the proposed model with other excellent models that have been widely used recently for medical image segmentation. Next, we perform an ablation study to validate the effectiveness of each block of the MLU-Net. Finally, it will be analyzed why our proposed method outperforms other methods and discussed what are the current shortcomings.

A. DATASETS

For our medical image segmentation experiments, we employed two distinct medical image datasets:

1) BRAIN TUMOR DATASET (MRI)

We chose the MRI data provided by Kaggle low grade glioma brain tumors. In contrast to high-grade gliomas, low-grade gliomas (LGG) typically exhibit less distinct tumor boundaries, presenting complex tissue structures on MRI images. These structures may intermingle with normal brain tissue, rendering precise tumor segmentation a challenging task. The dataset comprises data from 110 patients within The Cancer Genome Atlas (TCGA) [35], [36] collection, each accompanied by binary segmentation ground truth labels. Among these, 80 patient data were designated for training, 10 for validation, and the remaining 20 for testing. Prior to inputting the data into network model, preprocessing was conducted to standardize the images.

2) SKIN LESION DATASET (DERMATOSCOPE IMAGES)

We chose the publicly available ISIC2018 dataset, which includes a challenge with three image analysis tasks [37], [38]. Our focus in this study pertains to lesion segmentation. The Skin Lesion Segmentation Challenge comprises 2594 dermatoscope images and corresponding 2594 ground truth (GT) segmentation masks from 115 contributors, utilized for training. To assess model performance and prevent overfitting, we divided the 2594 dermatoscope images into training, validation, and test sets in a 7:1:2 ratio, resulting in 1814 training images, 260 validation images, and 520 test images. The validation set was employed for model training to optimize model parameters, which were then stored in model files. The test set served to evaluate various performance metrics of the model, and these metrics were used for model comparison against other state-of-the-art models.

B. DATA PRE-PROCESSING

Upon acquiring the datasets, we initiated preprocessing of the raw data prior to inputting it into the network. To ensure consistency and impartial performance evaluation across all experiments, we applied standardized preprocessing operations that were not model-specific but tailored to the inherent data characteristics.The inherent variability in color distribution within the dataset, arising from factors such as different patients, acquisition environments, and devices, prompted the need for these preprocessing steps. The process involved data partitioning, image resizing with cropping and padding to achieve uniform image dimensions, and color normalization.

As shown in the first row of histograms in Figure 7, the brain tumor data exhibited distinct peaks and modes among different images. For instance, histogram (a) illustrates a primary peak around 50 and a smaller one near 125, while histograms (b) and (c) exhibit peaks around 50. Furthermore, histogram (a) displays a broader peak span. Given the disparities in peak locations and spans, image normalization was implemented to harmonize color distributions across different data, rendering them more balanced, as presented

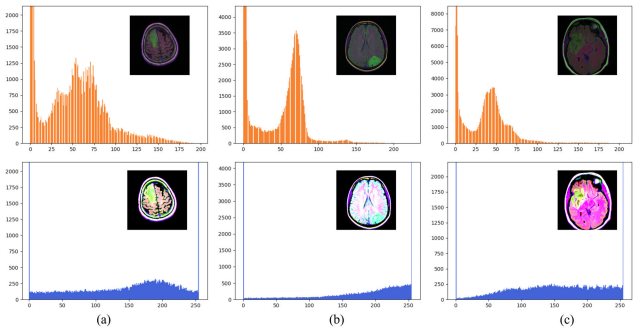


FIGURE 7. Histograms of some brain MRI samples before and after processing.

in the second row of histograms in figure B. Subsequently, indexing was conducted for the entire set of data from new patient samples. Following this series of operations, the preprocessed images were fed into the proposed MLU-Net and other network architectures for experimentation. Comparative analysis was performed on the experimental results obtained from different networks.

C. IMPLEMENTATION DETAILS

The experiments are performed using the Pytorch framework. The image size of the input network is set to 256 × 256. We set the batch size to 16 and the learning rate to 0.001 [39]. The Adam optimizer was chosen for training. We use these hyper-parameters uniformly in all experiments. During datasets training, we train the network for 300 epochs or until convergence. All experiments in this paper were conducted under identical computing specifications with an Intel i7 processor (3.6GHz), 32GB of RAM, and a 24GB graphics memory Nvidia RTX 3090 GPU on a 64-bit Windows 10 system.

D. EVALUATION METRICS

For quantitative performance assessment, several metrics were considered in the evaluation of comparative experiments. The Intersection over Union (IoU), also known as the Jaccard coefficient, serves as a metric to quantify the similarity between the segmented output and the ground truth in the context of image segmentation tasks [40]. Conversely, the Dice coefficient, an alternative measure, is employed for assessing the similarity between two sets [25]. The indicators are defined as shown in Equation (9) and Equation (10).

$$IoU = \frac{|Y_p \cap Y_t|}{|Y_p \cup Y_t|} = \frac{TP}{TP + FP + FN} \tag{9}$$

$$Dice = \frac{2\,|Y_p \cap Y_t|}{|Y_p| + |Y_t|} = \frac{2TP}{FP + 2TP + FN} \tag{10}$$

where, Y_p is predicted region and Y_t is *ground truth*(*GT*). *true positive* (*TP*) is defined as pixels that exist in both ground truth and the predicted segmentation region, and *true negative* (*TN*) represents pixels that are absent in both ground truth and the predicted segmentation region. In contrast, *false negative* (*FN*) corresponds to pixels that are present only in ground truth and *false positives* (*FP*) corresponds to pixels that are present only in predicted segmentation region.

E. ABLATION STUDIES

In this part, we delve into an extensive analysis of each module's effectiveness and its impact on the proposed MLU-Net through ablation experiments. The aim of these experiments is to elucidate the influence and efficacy of the proposed enhancements on the network. Our approach entails conducting thorough ablation experiments on each module within the multi-level network, MLU-Net, with the objective of validating the effects of the proposed improvements. We commence by dissecting the results of the ablation experiments conducted on various modules, subsequently honing in on the specific examination of the frequency domain layers, MLP layers, and varying network depths to further scrutinize their impacts.

1) OVERALL ABLATION EXPERIMENTS FOR THE METHODS INVOLVED

We selected U-Net as the baseline network and gradually added methods on it. The effect of the module on the network is verified by testing each method independently. This part of the experiment involves the network hopping connection method, the upsampling method, the downsampling method, the network depth, the MLP, and the comparison with the benchmark network, and the characteristics of each part can be clearly seen according to the experimental results in Table 1.

Within Table 1, the *Base* is baseline network defined by us as the U-Net. *ADD* denotes the utilization of element-wise addition during skip connections, whereas in its absence, multiple features along the channel dimension concatenation (Concat) are employed. *MDB* signifies the adoption of a multi-downsampling process, with alternatives involving the application of max-pooling methods. *SUB* designation pertains to the spectrum domain upsampling module, encompassing both spectrum domain upsampling and skip connections. Conversely, interpolation-based upsampling methods are employed when *SUB* is not applied. *LOW* is indicative of lower channel counts, specifically 16, 32, 128, 160, and 256, as opposed to the classical 64, 128, 256, 512, and 1024 network channel numbers. One of them, 160, is a number chosen from the middle of 128 and 256, which has been shown to provide good results [31]. *MLP* notation indicates the utilization of tokenized MLP-based methods in the fourth and fifth layers, with convolutional blocks employed for the preceding three layers. The experimental data underscore the influential impact of frequency representation methodologies on the overall model accuracy, with channel count reduction offering significant benefits for lightweight network enhancements. Each module exhibits distinct degrees of improvement for the network, making them versatile and applicable across various network architectures.

TABLE 1. Ablation Studies, where MLU-Net^2 denotes the MLU without MDB and SUB. MLU-Net^3 denotes the MLU with the MLP removed. The format of the evaluation index for IoU and Dice is "mean±std."

ADD	MDB	SUB	LOW	MLP	Networks	Params(M)	DICE	IoU
×	×	×	×	×	Base	34.52	85.71±0.23	88.97±0.46
✓	×	×	×	×	Base+ADD	31.38	87.32±0.34	90.54±0.48
×	✓	×	×	×	Base+MDB	34.52	88.22±0.78	91.38±0.64
×	×	✓	×	×	Base+SUB	34.52	86.81±0.55	90.02±0.57
×	×	×	✓	×	Base+LOW	0.93	87.77±0.37	91.02±0.51
×	×	×	×	✓	Base+MLP	33.51	87.88±0.49	91.12±0.46
×	✓	✓	×	✓	Base+MDB+SUB+MLP	34.52	86.99±0.60	90.12±0.61
×	✓	✓	✓	✓	Our Net-ADD	0.93	87.87±0.54	91.07±0.56
✓	×	×	✓	✓	MLU-Net^3	0.93	87.08±0.51	91.27±0.45
✓	✓	✓	✓	×	MLU-Net^2	**0.88**	88.18±0.44	91.06±0.56
✓	✓	✓	✓	✓	Our Net(MLU-Net^1)	**0.88**	**89.08±0.51**	**92.27±0.45**

2) THE ANALYSIS OF FREQUENCY DOMAIN LEVELS

In the multi-level U-Net network MLU-Net, a method for introducing frequency domain representations in the frequency domain hierarchy is employed. Within the frequency domain layers, we have designed multiple downsampling and upsampling modules dedicated to enhancing the learning capability of high-level semantic information relevant to segmentation targets. To elucidate their efficacy, we deliberately crafted an alternative version of MLU-Net, termed MLU-Net^2, that removed the multi-level downsampling and frequency domain upsampling methods and compared its performance with the complete module, MLU-Net^1. Experimental results in Table 1 unequivocally demonstrate the module's capacity to substantially enhance segmentation precision.

Furthermore, the investigation proceeds by dissecting the three-level network within the frequency domain hierarchy, delineating an encoding part and a decoding part. By leveraging pretrained model data, a 1 × 1 convolution is applied to all channel feature maps at each tier, culminating in channel-aggregated feature maps. A comparative analysis of feature maps during the encoding process in the first and second rows of Figure 8 reveals that in the MLU-Net^1, where the frequency domain method is introduced, the overall information content of the images is consistently more pronounced, with a richer information reservoir throughout the encoding process compared to the traditional approach. This enhancement can be attributed to the effective preservation of low-frequency spectral information within the encoding structure's multiple downsampling modules, encouraging the network to capture more contours and structural information during the learning process, thus facilitating more effective feature information capture and superior segmentation outcomes.

Similarly, within the decoding segment of the network, the employment of frequency domain up-sampling modules distinctly emphasizes the retention of low-frequency high-level semantic features and high-frequency fine-grained details within the spectral domain. This strategy effectively restores high-level semantic features and high resolution, ultimately yielding high-precision predictions of pathological

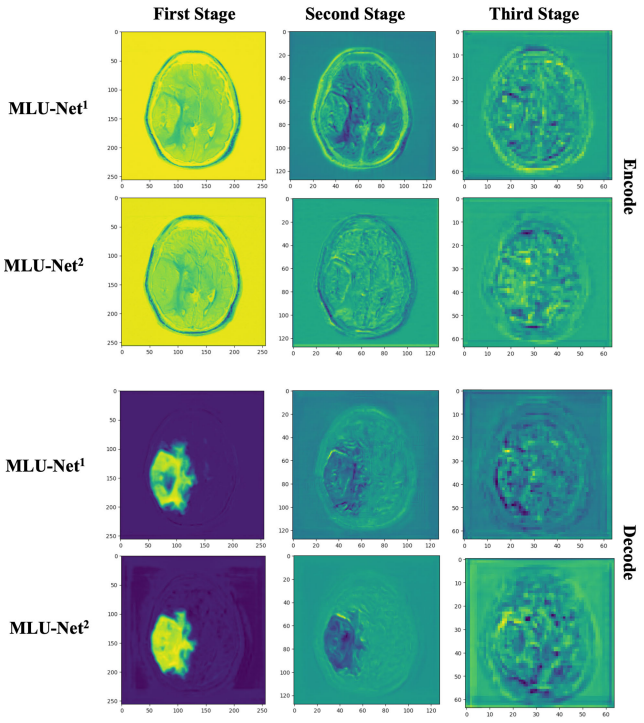


FIGURE 8. The channel merged feature maps are obtained by 1 × 1 convolution of all the channel feature maps in the three-layer network at the frequency domain level. It is specifically divided into encoding part and decoding part. Compare MLU-Net^1 with MLU-Net^2 after removing the frequency domain representation.

segmentation images. The visual comparison of images in the third and fourth rows of Figure 8 clearly underscores the enrichment of information in the decoding process while accurately capturing critical regions of brain tumor pathology.

3) THE ANALYSIS OF MLP LEVELS

Subsequently, we conducted experimental analysis on the multilayer perceptron (MLP) levels within the MLU-Net architecture. The integration of MLP enriches the network's capabilities by extending its capacity to capture global segmentation target information, beyond the realm of local

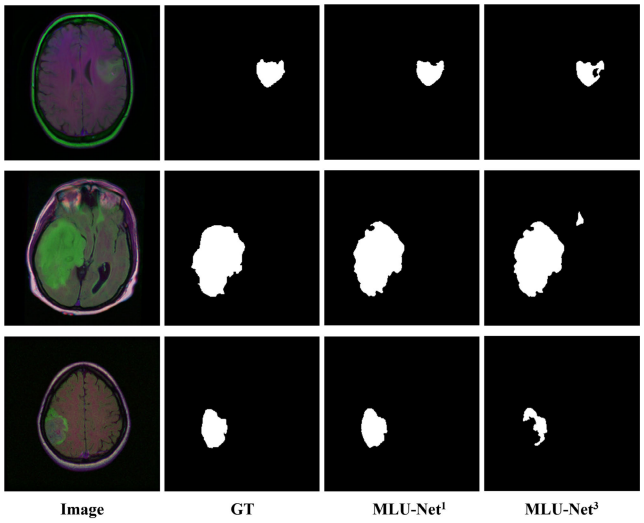


FIGURE 9. Comparison of the overall structure of MLU-Net[1] with MLU-Net[3] that removes the MLP method in the partial segmentation prediction target of brain tumor images.

information. This enhancement augments the network's ability to glean latent information, thereby improving the precision of the entire lesion segmentation process.

To further illustrate this, we conducted comparative experiments between the complete MLU-Net[1] and a modified MLU-Net[3], from which the MLP-based methods is removed. Owing to its heightened emphasis on global information, MLU-Net exhibits a more pronounced focus on the overall structural information when segmenting lesion targets. This effectively mitigates issues such as segmentation region omissions and extraneous details in regions beyond the segmentation boundary. As demonstrated in Figure 9, when confronted with images of intricate or ambiguous structures, networks employing the MLP-based approach are better equipped to accurately delineate the expected results in the context of lesion segmentation.

4) ANALYSIS OF THE NUMBER OF NETWORK CHANNELS

Furthermore, our investigation on the impact of channel count on various aspects of network performance within the MLU-Net framework is presented in Table 2. Here, *LOW* encompasses channel counts of 16, 32, 128, 160, and 256, reflecting our selected channel numbers. *MID* encompasses 32, 64, 128, 256, and 512, and *HIGH* represents 64, 128, 256, 512, and 1024. The segmentation experiments on brain tumors illustrate that the reduction in channel count not only significantly reduces network computational overhead but also enhances performance to some extent without compromising network efficacy.

F. EXPERIMENTS ON BRAIN TUMORS DATASETS

1) COMPARISON OF SEGMENTATION EFFECTS

We compared the performance of MLU-Net with widely used medical image segmentationframeworks, which are

TABLE 2. Comparison experiments with different number of channels, the format of the evaluation index for IoU and Dice is "mean±std."

Networks	Params(M)	FLOPs(G)	IoU	Dice
MLU-Net-HIGH	31.39	41.26	86.57±0.54	89.85±0.67
MLU-Net-MID	7.85	10.36	86.95±0.49	90.02±0.56
MLU-Net-LOW	**0.88**	**1.07**	**89.08±0.31**	**92.27±0.35**

TABLE 3. Comparison of metrics for brain tumor segmentation results, the format of the evaluation index for IoU and Dice is "mean±std."

Networks	Params(M)	FLOPs(G)	IoU	Dice
U-Net	34.52	65.43	85.71±0.23	88.97±0.46
ResUNet50	72.22	43.92	85.53±0.31	88.76±0.52
U-Net++	36.63	138.59	86.10±0.21	89.43±0.44
R2Net	39.09	152.82	79.75±0.74	83.78±0.44
R2AttNet	39.44	149.06	81.97±0.23	85.36±0.24
UNeXt	9.48	**0.57**	86.19±0.28	89.74±0.23
Our Net	**0.88**	1.07	**89.08±0.31**	**92.27±0.35**

U-Net [11], U-Net++ [12], [13] and ResUNet [24] by using of convolutional methods, and R2Net [20] and R2AttNet by using of transformer baselines. Dice coefficient and Intersection over Union (IoU) score constitute critical components in the evaluation of segmentation performance metrics, taking into consideration the computational complexity as reflected by the number of model parameters. Our experiment results shown in Table 3, clearly demonstrate that MLU-Net outperforms all the base networks in terms of segmentation quality and computational efficiency. The segmentation effect of different networks on some brain tumor images is demonstrated in Figure 10.

However, the most striking point here is the disparity in the number of parameters, where MLU-Net and UNeXt [31] have substantially smaller computations than the other networks, because these two networks do not use a large number of network channels and complex processing modules. In particular, we note that UNeXt has only 9.48M parameters, while MLU-Net has 0.88M parameters. The MLU-Net network is heavily optimized for light weight, while ensuring better network performance. In Figure 10, we observe that MLU-Net achieves excellent results across various types of brain tumor samples.

2) COMPARISON OF PARAMETERS AND FLOPS

We show the IoU indices and parameters of the different network models along with the scatter plot of computational FLOPs, the closer the location of the points in the figure to the upper left the more desirable the effect is. It is clear from Figure 11 that proposed model performs well in terms of performance and complexity and does not lose the network performance due to the lightweighting of the network.

G. EXPERIMENTS ON SKIN LESION DATASETS

Furthermore, to validate the robustness of the model, we transitioned to an entirely different type of medical image data. Segmentation experiments were conducted on lesion data obtained from skin imaging using dermatoscopes. Unlike brain tumor MRI images, skin data is captured through a lens,

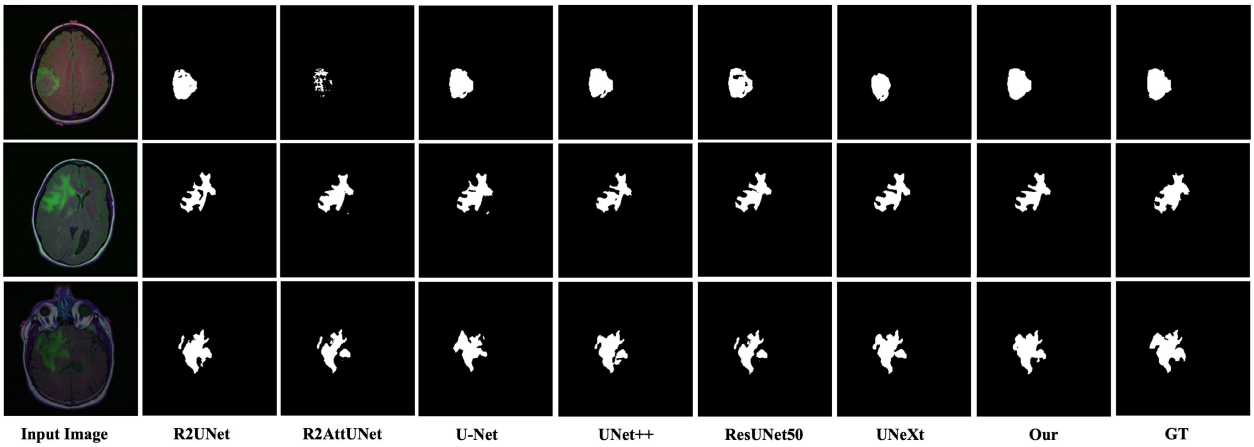
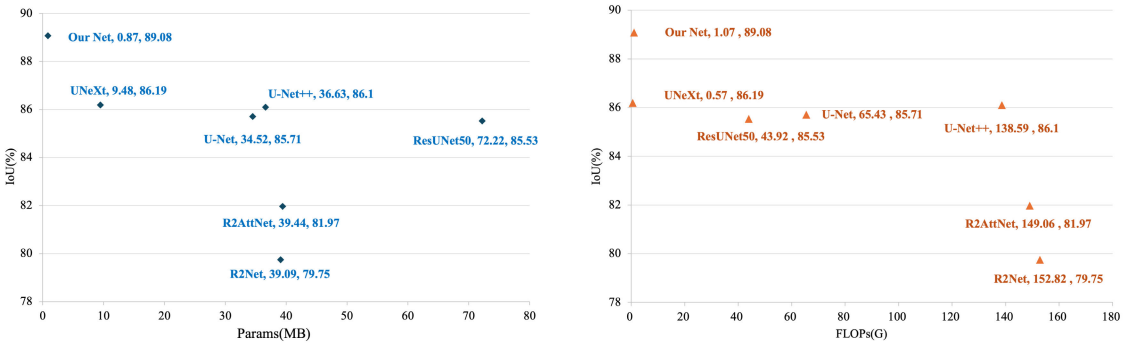


FIGURE 10. Some effects of different networks in brain tumor segmentation.



(a) IoU and Parameters (b) IoU and FLOPs

FIGURE 11. Scatter plot of different networks IoU compared to Parameters and FLOPs respectively (the closer the points are to the top left indicates better results).

resulting in rich and diverse color variations with varying lesion types and structures. In this experiment, UNeXt [31] demonstrated competitive performance. Benefiting from the feature-preserving capability of the frequency representation method and the computational efficiency of MLP, our network maintained a competitive edge in segmentation accuracy under a lightweight structure compared to other networks, as shown in Table 4. It is noteworthy that the proposed network did not exhibit as pronounced an advantage in the skin lesion dataset as observed in the brain tumor data. This analysis may be attributed to the relatively small size of the dataset, highlighting the network's capability more prominently. The segmentation results are illustrated in Figure 12.

TABLE 4. Results of skin lesion segmentation on different networks, the format of the evaluation index for IoU and Dice is "mean±std."

Networks	IoU	Dice
U-Net	78.03±0.56	86.03±0.68
ResUNet50	78.32±0.32	86.20±0.46
U-Net++	77.41±0.46	85.58±0.54
R2Net	70.87±0.76	80.53±0.75
R2AttNet	72.17±0.28	81.79±0.12
UNeXt	78.54±0.54	86.56±0.51
Our Net	**79.36±0.42**	**87.05±0.49**

V. DISCUSSION AND FUTURE WORK

In this work, we introduce a multi-level segmentation network, MLU-Net, based on the U-Net architecture. The overarching structure of MLU-Net comprises three frequency domain layers and two multilayer perceptron (MLP) layers, delivering exemplary segmentation accuracy while maintaining network lightweight attributes. Within the frequency domain layers, we incorporate multiple downsampling and frequency domain upsampling modules that are specifically tailored for downsampling and upsampling processes. These modules enhance the network's capacity to learn high-level semantic information relevant to segmentation targets. Furthermore, the modules built on this method can be conveniently integrated to replace existing deep learning modules. The application of MLP layers involves tokenized MLP modules for learning latent feature information, contributing to the network's capability to supplement global information of the segmentation target, beyond its proficiency

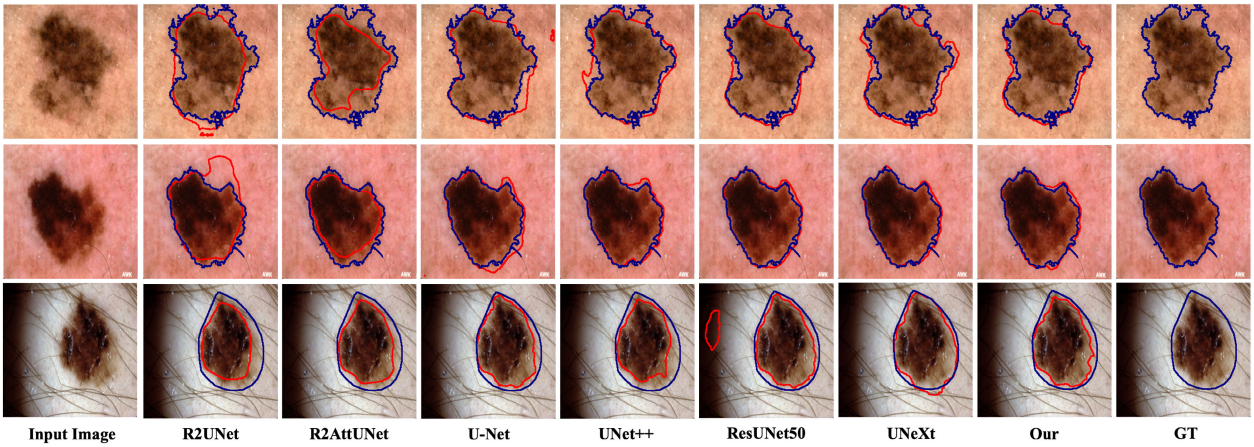


FIGURE 12. Effectiveness of different networks in skin lesion segmentation. The blue line is the ground truth and the red line is the prediction.

in local feature learning. This enhancement serves to augment the network's ability to acquire richer feature information, thereby elevating the precision of the entire lesion target segmentation process.

We conducted comprehensive ablation experiments on brain tumor datasets, revealing that both the frequency domain and MLP layers within MLU-Net play pivotal roles in enhancing the network segmentation model. Furthermore, we extended our experiments to skin lesion segmentation tasks, with results showcasing the exceptional performance of MLU-Net across distinct datasets. In the brain tumor segmentation experiment, our network has better experimental results than U-Net by use of only 1/39 of the number of parameters and 1/61 of the computation under the same preprocessing condition. The experiments mean that the proposed lightweight approach could be applied to a wider range of medical scenarios in the future.

MLU-Net excels in its feature information awareness and lightweight characteristics. However, it exhibits diminished efficacy when confronted with large-scale data tasks. For instance, its advantages in skin lesion segmentation are somewhat less pronounced compared to brain tumor segmentation, highlighting the challenge of reconciling network lightweight attributes with task scale–a conundrum faced by many researchers. In this study, our research focused exclusively on images of brain tumors and skin lesions, and the scalability of the proposed methodology is a subject of inquiry.

In forthcoming research, we intend to delve deeper into the spatial relationships among image pixels and their neighboring points. This is pivotal for understanding latent patterns within images and optimizing image preprocessing. Additionally, we plan to conduct a more profound exploration of frequency domain characteristics and introduce attention mechanisms to facilitate a more in-depth investigation into medical image segmentation tasks. Such research endeavors are poised to make valuable contributions to the development of critical medical assistive tools for future healthcare applications.

The efficient and rapid segmentation capabilities of lightweight medical image segmentation networks contribute to enhanced diagnostic efficiency for healthcare professionals. Particularly, in real-time monitoring and surgical assistance, these networks offer precise information to facilitate surgical procedures. Such networks are well-suited for mobile medical applications, facilitating remote diagnostics and image analysis in mobile environments, with particular benefits for resource-constrained regions. The lightweight nature of these models reduces data processing requirements, thereby aiding in improving data privacy and security, and mitigating the risks associated with patient information transmission. In terms of cost-effectiveness, the adoption of lightweight models reduces healthcare equipment costs, presenting economic advantages for resource-limited healthcare institutions. In summary, lightweight medical image segmentation networks play a pivotal role in advancing the overall efficiency, cost-effectiveness, and security of medical image processing.

VI. CONCLUSION

In this study, we propose a multi-layer lightweight U-Net network named MLU-Net, which focuses on the segmentation of structurally abnormal regions of tumors and lesions in medical images. MLU-Net adopts a lightweight network structure including frequency domain layers and layers based on the MLP method, which significantly reduces the network parameters and computational complexity, and at the same time, it can efficiently extract tumor and lesion regions in medical images. At the same time, it can effectively extract the low-frequency contour information and overall structure information of the lesion region in medical images, realizing fast training and accurate segmentation of the model. This makes medical image segmentation more applicable to various clinical scenarios.

REFERENCES

[1] D. Kollias, Y. Vlaxos, M. Seferis, I. Kollia, L. Sukissian, J. Wingate, and S. Kollias, "Transparent adaptation in deep medical image diagnosis," in *Proc. Int. Workshop Found. Trustworthy AI Integrating Learn., Optim. Reasoning*, 2020, pp. 251–267.

[2] D. Kollias, A. Arsenos, and S. Kollias, "AI-MIA: COVID-19 detection and severity analysis through medical imaging," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 677–690.

[3] D. Kollias, A. Arsenos, L. Soukissian, and S. Kollias, "MIA-COV19D: COVID-19 detection through 3-D chest CT image analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 537–544.

[4] J. Wingate, I. Kollia, L. Bidaut, and S. Kollias, "Unified deep learning approach for prediction of Parkinson's disease," *IET Image Process.*, vol. 14, no. 10, pp. 1980–1989, Aug. 2020.

[5] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, and H. J. Aerts, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA, Cancer J. Clinicians*, vol. 69, no. 2, pp. 127–157, 2019.

[6] D. D. Patil and S. G. Deore, "Medical image segmentation: A review," *Int. J. Comput. Sci. Mobile Comput.*, vol. 2, no. 1, pp. 22–27, 2013.

[7] H. Seo, M. Badiei Khuzani, V. Vasudevan, C. Huang, H. Ren, R. Xiao, X. Jia, and L. Xing, "Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications," *Med. Phys.*, vol. 47, no. 5, pp. e148–e167, May 2020.

[8] P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, and W. Enbeyle, "Deep neural networks for medical image segmentation," *J. Healthcare Eng.*, vol. 2022, pp. 1–15, Jul. 2022.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[10] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Springer, 2015, pp. 234–241.

[12] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. 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*. Springer, 2018, pp. 3–11.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[14] J. Tian, D. Dong, Z. Liu, and J. Wei, "Introduction," in *Radiomics and Its Clinical Application*. Academic, 2021, pp. 1–18. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128181010000045

[15] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Spectral representations for convolutional neural networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (ACM)*, 2015, pp. 2449–2457. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/536a76f94cf7535158f66cfbd4b113b6-Paper.pdf

[16] H. Pratt, B. Williams, F. Coenen, and Y. Zheng, "FCNN: Fourier convolutional neural networks," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 786–798.

[17] S. O. Ayat, M. Khalil-Hani, A. A.-H. Ab Rahman, and H. Abdellatef, "Spectral-based convolutional neural network without multiple spatial-frequency domain switchings," *Neurocomputing*, vol. 364, pp. 152–167, Oct. 2019.

[18] M. Pezhman Pour and H. Seker, "Transform domain representation-driven convolutional neural networks for skin lesion segmentation," *Expert Syst. Appl.*, vol. 144, Apr. 2020, Art. no. 113129.

[19] A. Dziedzic, J. Paparrizos, S. Krishnan, A. Elmore, and M. Franklin, "Band-limited training and inference for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 1745–1754.

[20] M. Zahangir Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R^2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[21] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with Densley connected convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 406–415.

[22] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, vol. 2018, pp. 269–284.

[23] X. Li, H. Chen, X. Qi, Q. Dou, C. Fu, and P. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[24] S. Vesal, N. Ravikumar, and A. Maier, "A 2D dilated residual U-Net for multi-organ segmentation in thoracic CT," 2019, *arXiv:1905.07710*.

[25] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[26] Z. Zhang, C. Wu, S. Coleman, and D. Kerr, "DENSE-inception U-Net for medical image segmentation," *Comput. Methods Programs Biomed.*, vol. 192, Aug. 2020, Art. no. 105395.

[27] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[28] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, and M. Lucic, "MLP-mixer: An all-MLP architecture for vision," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[30] D. Lian, Z. Yu, X. Sun, and S. Gao, "AS-MLP: An axial shifted MLP architecture for vision," 2021, *arXiv:2107.08391*.

[31] J. Valanarasu and V. M. Patel, "UNeXt: MLP-based rapid medical image segmentation network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2022, pp. 23–33.

[32] G. Kreiman, *Biological and Computer Vision*. Cambridge Univ. Press, 2021.

[33] X. Tang, J. Peng, B. Zhong, J. Li, and Z. Yan, "Introducing frequency representation into convolution neural networks for medical image segmentation via twin-kernel Fourier convolution," *Comput. Methods Programs Biomed.*, vol. 205, Jun. 2021, Art. no. 106110. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721001851

[34] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[35] R. Mclendon, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–1068, 7216.

[36] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.

[37] N. Codella, V. Rotemberg, P. Tschandl, M. Emre Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[38] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, Aug. 2018.

[39] Y. Li, Q. Zhang, and S. W. Yoon, "Gaussian process regression-based learning rate optimization in convolutional neural networks for medical images classification," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115357. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421007855

[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

LIPING FENG received the B.S. and M.S. degrees in computer application from Shanxi University, in 1999 and 2006, respectively, and the Ph.D. degree in computer application from Chongqing University, in 2013. She is currently a Professor with the Department of Computer Science, Xinzhou Teachers University. Her research interests include distributed optimization, network security, and dynamical modeling.

KEPENG WU is currently pursuing the master's degree with the Chongqing University of Science and Technology, Chongqing, China. His research interests include image segmentation, image super-resolution, image refinement reconstruction, medical image noise reduction, and neural network parameter optimization. His current research interests include MRI medical image segmentation and medical auxiliary diagnosis.

ZIYI PEI is currently pursuing the degree in material forming and control engineering with the Chongqing University of Arts and Sciences. His current research interests include artificial intelligence and deep learning.

TENGFEI WENG received the B.S. and M.S. degrees in chemistry and chemical engineering from Chongqing University, China, in 2007 and 2010, respectively. She is currently with the Chongqing University of Science and Technology. Her current research interests include artificial intelligence and neural networks.

QI HAN (Member, IEEE) received the B.S. degree in computer science and technology from Shandong University, China, in 2005, and the M.S. and Ph.D. degrees from Chongqing University, China, in 2009 and 2012, respectively. He is currently an Associate Professor with the Chongqing University of Science and Technology. His current research interests include artificial intelligence, system optimization, neural networks, and chaos control.

LUN MENG is currently pursuing the master's degree with the Chongqing University of Science and Technology. His current research interests include computer vision and semantic segmentation.

XIN QIAN received the bachelor's degree in computer science from Jinggangshan University, China, in 2021. He is currently pursuing the master's degree with the Chongqing University of Science and Technology. His research interests include machine learning and medical image analysis.

HONGXIANG XU received the B.S. degree from Sanjiang University, Jiangsu, China, in 2020. He is currently pursuing the master's degree with the Chongqing University of Science and Technology. His research interests include deep learning, brain computer interface, and medical image analysis.

ZICHENG QIU was born in Wuhan, Hubei, China, in 1983. He received the B.S. degree in optoelectronic engineering from the Huazhong University of Science and Technology, Wuhan, in 2005, and the Ph.D. degree in optical engineering (advanced lithography technologies) from the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 2010. From 2010 to 2011, he was a Software Engineer with Synopsys. From 2011 to 2015, he was an Assistant Professor with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Science, Chongqing, China. From 2015 to 2020, he was an Associate Professor with the College of Information Engineering, Tarim University, Alar, Xinjiang, China. He is currently an Associate Professor with the College of Intelligent Technology of Engineering. He is the author of more than ten articles and three inventions. His research interests include AI, intelligent speech technologies, and security of AI.

ZHONG LI is currently a Distinguished Scholar and a Researcher of computer science and intelligent technology. With a career spanning from July 2006 to December 2022, as a Lecturer with the School of Intelligent Technology and Engineering, Chongqing University of Science and Technology, he has been advanced to the role of an Associate Professor, since January 2023. He also holds the position of an associate professor with the School of Intelligent Technology and Engineering, Chongqing University of Science and Technology. He has published numerous research articles in artificial intelligence and control systems.

YUAN TIAN received the B.S. degree in computer science and technology from Chongqing Normal University, Chongqing, China, in 2009, the M.S. degree in computer application technology from Chongqing University, Chongqing, in 2012, and the Ph.D. degree in computational intelligence and information processing from Southwest University, Chongqing, in 2020. She was with the Chongqing University of Science and Technology, Chongqing. Her current research interests include impulsive systems, discontinuous dynamical systems, and multi-agent systems.

GUANZHONG LIANG received the bachelor's degree in clinical medicine from the Chuannorth Medical College, in 2009, and the master's degree in oncology from Shanxi Medical University, in 2012. Since July 2012, he has been dedicated to medical oncology with Affiliated Tumor Hospital. He is currently a Medical Professional and also the Chief Physician with Affiliated Tumor Hospital, Chongqing University Cancer Hospital. His contributions to research projects, including the study of mechanisms to enhance the effectiveness of PD-1 inhibitors in non-small cell lung cancer, have significantly impacted cancer treatment. He has also shared his expertise through research papers and presentations.

YAOJUN HAO received the Ph.D. degree in computer science and technology from Yanshan University, in 2020. He is currently a Full Professor with Xinzhou Normal University, China. His research interests include information security and recommender systems.