Scilight

Sensors and AI
https://www.sciltp.com/journals/sai

Review

Advancements and Challenges in Medical Image Segmentation: A Comprehensive Survey

Guanqiu Qi [1,*], Zhiqin Zhu [2], Ke Li [3] and Han Xiao [3]

[1] Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222, USA
[2] College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[3] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
* Correspondence: qig@buffalostate.edu

How To Cite: Qi, G.; Zhu, Z.; Li, K.; et al. Advancements and Challenges in Medical Image Segmentation: A Comprehensive Survey. *Sensors and AI* **2025**, *1*(1), 3–29.

Received: 16 December 2024
Revised: 17 February 2025
Accepted: 18 February 2025
Published: 11 March 2025

Abstract: Medical image segmentation is a fundamental task in the field of medical imaging, enabling the accurate identification and delineation of structures such as organs, tissues, and lesions within medical images. These segmented regions are essential for diagnostic purposes, treatment planning, and disease monitoring. Over the years, medical image segmentation has evolved significantly, driven by advances in imaging technologies and computational techniques. Traditional methods, such as thresholding, region-growing, and active contours, have been supplemented and, in some cases, replaced by more sophisticated machine learning (ML) and deep learning (DL) approaches. Convolutional neural networks (CNNs) and their variants, including U-Net and Transformer-based models, have shown remarkable success in automating and improving segmentation tasks. This survey paper provides a comprehensive review of the various segmentation techniques, categorizing them into classical and deep learning-based methods. It discusses the strengths, limitations, and challenges of each approach, including issues related to data quality, class imbalance, and the generalizability of models. Furthermore, the paper highlights recent advancements in the field, emerging trends, and future directions for further enhancing segmentation accuracy, robustness, and efficiency in clinical applications. This work aims to serve as a valuable resource for researchers and clinicians looking to understand the current state of medical image segmentation and its potential future developments.

Keywords: medical image segmentation; encoding-decoding; deep learning; machine learning; convolutional neural networks; multi-modality

1. Introduction

Medical image segmentation is a crucial task in the field of medical imaging, enabling the extraction of meaningful structures or regions of interest (ROI) from complex and often noisy medical images [1]. It plays a pivotal role in a wide array of clinical applications, including but not limited to disease diagnosis, surgical planning, treatment monitoring, and radiotherapy [2, 3]. Segmentation algorithms aim to delineate boundaries of anatomical structures such as organs, tissues, tumors, and blood vessels, which are essential for accurate diagnosis and therapeutic interventions [4].

Multi-modality image information refers to images that are captured using different techniques or sensors, providing complementary information [5–7]. The advancements in multi-modality imaging technologies, such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and ultrasound, have significantly improved the resolution and quality of medical images [8]. CT gives detailed images of bone structures. MRI provides high-resolution images of soft tissues. PET shows metabolic processes in the body. Ultrasound provides real-time images and is useful for examining soft tissues and organs. By combining these

images, healthcare providers can gain a more complete and detailed understanding of the patient's condition. However, the complexity of medical images, often characterized by varying contrast, noise, low signal-to-noise ratios, and anatomical variability, poses significant challenges for traditional image processing techniques [9, 10]. As a result, medical image segmentation has become a focus of intense research, especially with the rise of machine learning (ML) and deep learning (DL) techniques, which offer robust and automated solutions for segmentation tasks [11–13].

This survey paper aims to provide a comprehensive overview of the state-of-the-art techniques and methodologies in medical image segmentation. Classical approaches are reviewed, including thresholding, region growing, and active contour models, as well as more recent deep learning-based approaches, such as convolutional neural networks (CNNs) and their variants. Additionally, the challenges in medical image segmentation are explored, such as handling small datasets, managing class imbalances, and ensuring generalizability across diverse populations and imaging modalities.

The goal of this paper is to provide researchers and practitioners with an in-depth understanding of the various methods, their strengths and limitations, and future directions in the evolving landscape of medical image segmentation. The rest of this paper is organized as follows. Section 2 discusses medical image segmentation; Section 3 explores brain tumor segmentation; Section 4 compares existing experimental results; and Section 5 concludes this paper and points out future directions.

2. The Overview of Medical Image Segmentation

As an important sub-branch in the field of medical image processing and analysis, the automatic segmentation techniques are constantly advancing. Related techniques have made considerable progress in the past 30 years. The following paragraphs classify and discuss the common medical image segmentation solutions.

2.1. Traditional Medical Image Segmentation Methods and Related Progress

Traditional medical image segmentation methods primarily rely on machine learning, with several standard techniques including threshold segmentation, region-based segmentation, and edge-based segmentation methods [14]. The threshold segmentation algorithm is centered on selecting an appropriate threshold value, which allows for binary segmentation of medical images to distinguish between tumors and normal tissues [15]. However, this approach is highly sensitive to noise and image variations, potentially leading to inaccurate segmentation results. Region-growing and merging methods are common in sequence-based region correlation techniques [16], where the results of earlier steps influence the subsequent segmentation stages. The accuracy of these methods can be impacted by noise and the quality of initial seed points. Edge-based segmentation methods detect boundaries based on grayscale differences, enabling the segmentation of image frames [17]. However, these methods can produce inaccurate results when dealing with discontinuous boundaries or complex shapes. Finally, segmentation methods based on specific theories, such as those utilizing cluster analysis, fuzzy set theory, or wavelet transforms, propose new approaches rooted in various disciplines [18], offering additional options for medical image segmentation.

In recent years, the rapid advancements in deep learning have led to widespread attention on recognition techniques based on artificial neural networks, particularly for image segmentation. The neural network-based segmentation approach begins by training a multilayer perceptron to derive a linear decision function. This model then classifies the pixels essential for achieving accurate image segmentation. Neural networks, with their extensive connectivity, are well-suited to incorporate spatial information, helping to address challenges such as noise and uneven distribution in the input images. Due to their strong learning capabilities and adaptability, these networks are effective in handling multimodal data and utilizing boundary information, making them highly effective for tasks like brain tumor segmentation. As a result, deep learning-based methods have become widely adopted and offer superior overall performance.

2.2. Region-Based, Statistics-Based, and Fuzzy Theory-Based Segmentation Methods

The core concept of region-based segmentation methods is the similarity of features within a region, where the internal features of the same object are similar, while features between different objects are not continuous. To implement this idea, various approaches, such as thresholding, region growing [19], and random field methods [20], are often used to segment medical images. A statistical approach [21] focuses on the fact that the gray value of pixels at the edges of regions tends to change significantly, making segmentation possible by detecting changes in edge pixels across different areas. Image segmentation is typically a poorly structured problem, and fuzzy set theory, which is well-suited to handle such issues, is also employed for medical image segmentation. This includes methods such as fuzzy clustering segmentation [22] and fuzzy connection degree segmentation [23]. While these

methods have demonstrated varying levels of effectiveness, they often require manual intervention for effective feature selection and have certain limitations.

2.3. Segmentation Methods Based on Statistical Features

Statistical features, particularly low-dimensional texture features, play a crucial role in enhancing the performance of semantic segmentation. Many existing approaches leverage texture information derived from statistical features. For example, Simonyan et al. [24] utilized Fisher vector layers to enhance features through handcrafted techniques. Wang et al. [25] introduced learnable histograms for semantic segmentation and object detection. Additionally, Zhu et al. [26] proposed a texture enhancement module and a pyramid texture extraction module to extract image texture features, thereby improving the effectiveness of semantic segmentation.

2.4. Semantic-Based Segmentation Methods

Over the past decade, new techniques such as Convolutional Neural Networks have gained significant attention due to their ability to automatically extract pixel-level features and deliver superior image segmentation performance. The introduction of Fully Convolutional Networks (FCNs) [27] marked the beginning of the neural network era for semantic segmentation, followed by the development of U-Net [28], a neural network architecture particularly well-suited for medical image segmentation. The original U-Net architecture utilizes an encoder-decoder structure to extract both high- and low-level contextual features, while skip connections preserve spatial information, enabling effective fusion of deep and shallow features. Building on this foundational design, researchers have proposed various enhancements. U-Net++ [29], for example, introduces denser skip connections, as illustrated in Figure 1. The key innovation of U-Net++ lies in its nested dense skip pathways, formulated as follows:

$$x^{i,j} = \begin{cases} H\left(x^{i-1,j}\right), & j = 0 \\ H\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, U(x^{i+1,j-1})\right]\right), & j > 0 \end{cases} \tag{1}$$

where $x^{i,j}$ is the output at node $X^{i,j}$, with i as the encoder's downsampling level and j as the index in the skip pathway; $H(\cdot)$ is a convolution with activation, $U(\cdot)$ is upsampling, and $[\cdot]$ denotes concatenation. For $j = 0$, $x^{i,j}$ is computed from $x^{i-1,j}$. For $j > 0$, $x^{i,j}$ uses outputs from previous nodes $x^{i,0}, \ldots, x^{i,j-1}$ and the upsampled feature from $x^{i+1,j-1}$. This design aligns the features from both the encoder and decoder, enhancing segmentation performance.

ResU-Net [30] and DenseU-Net [31] modified U-Net by replacing each sub-module with residual and dense connections, respectively. Attention mechanisms were also incorporated into U-Net [32]. Before merging the encoder features with the corresponding decoder features, an attention module was used to adjust the encoder's output. Attention U-Net [33] introduced attention mechanisms into the skip connections. Additionally, U-Net has been extended to 3D for three-dimensional image segmentation [34]. While these U-Net variants have been designed to address various issues and improve segmentation accuracy, they have remained within the full convolution framework, overlooking the significant long-range dependencies between pixels, which presents a limitation to further advancements.

In recent years, the Transformer architecture has showcased exceptional global modeling capabilities across various computer vision tasks, including image segmentation. Transformer-based methods segment the input image into patches and apply self-attention mechanisms to these patches. Swin Transformer [35] improves on this approach by utilizing shifted windows to calculate attention across different feature map layers. Similarly, Vision Transformer (ViT) [36] has proven its strong modeling abilities in computer vision tasks, splitting the input image into patches and performing self-attention operations on them. While ViT focuses on self-attention across the entire image, Li [37] utilizes Restormer and Transformer layers for feature extraction, employing bidirectional stepwise feature alignment (BSFA) to predict deformation fields. This approach helps align unaligned image features, minimizing modality discrepancies and ensuring accurate multimodal image fusion. MedT [38] introduces enhanced gated self-attention and applies Transformer-based techniques to medical image segmentation tasks, as illustrated in Figure 2. The key formula for Gated Axial Attention is defined as:

$$y_{ij} = \sum_{w=1}^{W} \text{softmax}\left(q_{ij}^T k_{iw} + G_Q q_{ij}^T r_{iw}^q + G_K k_{iw}^T r_{iw}^k\right)\left(G_{V1} v_{iw} + G_{V2} r_{iw}^v\right) \tag{2}$$

where y_{ij} is the output at position (i, j), q_{ij}, k_{iw}, v_{iw} are the query, key, and value vectors, and r_{iw}^q, r_{iw}^k, r_{iw}^v are the relative positional encodings. G_Q, G_K, G_{V1}, G_{V2} are learnable gating parameters controlling the influence of positional encodings. The softmax normalizes attention weights over width W. This design enables MedT

to regulate the contributions of positional encoding, enhancing segmentation performance, particularly on small medical image datasets.

Recent approaches have attempted to combine the strengths of CNNs and Transformers by integrating both architectures into a novel backbone network. The CMT (Convolutional neural networks Meet Transformers) [39] block, for example, merges a depthwise convolution-based local perception unit with a lightweight Transformer module. CoAtNet [40] integrates these two structures using MBConv (Mobile inverted Bottleneck Convolution) and relative self-attention, exploring their potential fusion. The SETR (SEgmentation TRansformer) model proposed by Zheng et al. [41] completely replaces the CNN encoder with a Transformer encoder, offering the first verification of Transformer structures in image segmentation tasks. However, due to the absence of convolutional data's spatial inductive bias, local information modeling remains insufficient. The TransU-Net model introduced by Chen et al. [42] proposes a two-stage encoder structure (CNN to Transformer), integrating both architectures for medical image segmentation. However, this two-stage encoding approach performs a secondary extraction of image features, neglecting the equal importance of both local and global features, leading to insufficient integration of the two. Additionally, methods like TransBTS (Transformer-based Brain Tumor Segmentation) [43] combine Transformers and U-Net for 3D brain tumor segmentation. While these approaches have improved segmentation performance in both natural and medical images, they have not fully leveraged the strengths of both CNNs and Transformers. Furthermore, performance remains limited, particularly for small-scale datasets.

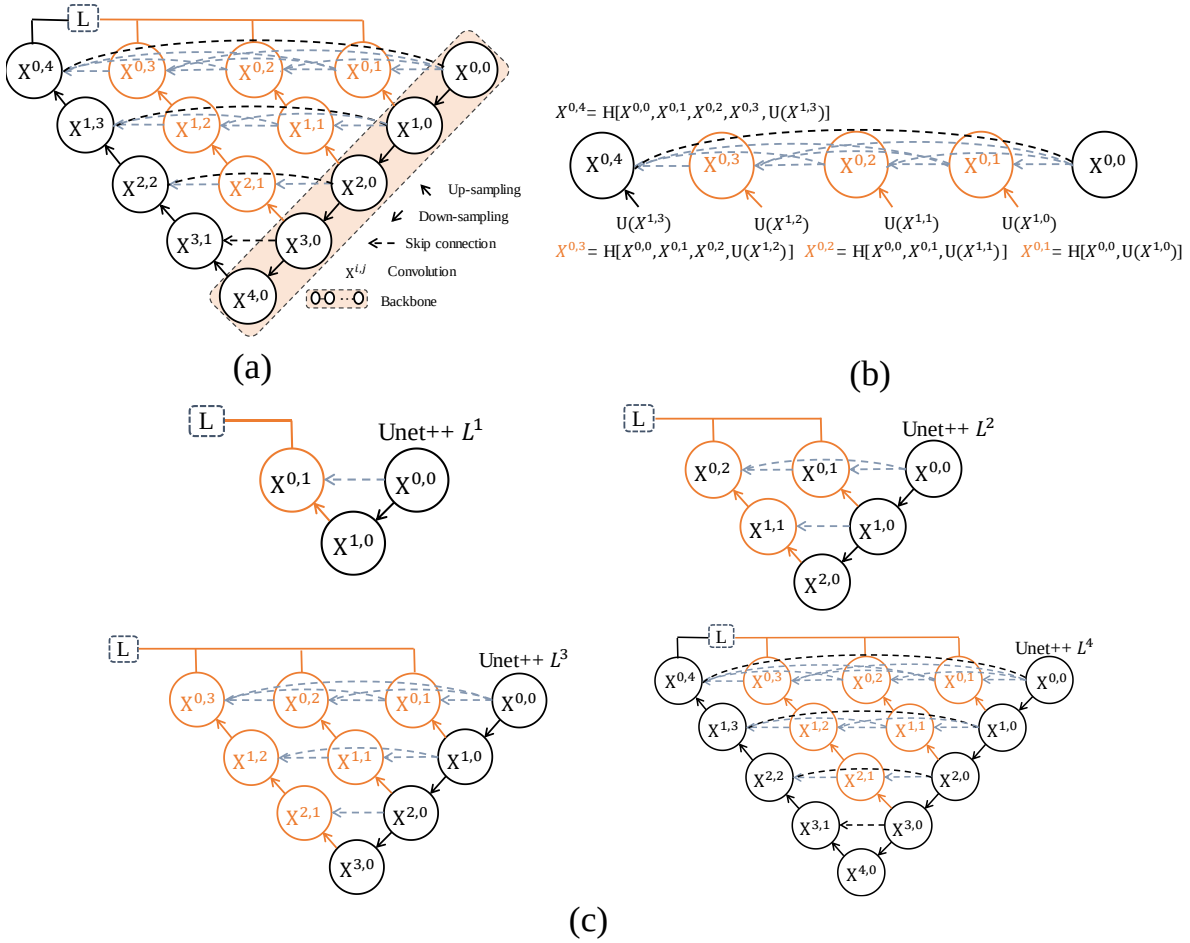

Figure 1. A high-level overview of UNet++. (**a**) UNet++ consists of an encoder and decoder that are connected through a series of nested dense convolutional blocks. The main idea behind UNet++ is to bridge the semantic gap between the feature maps of the encoder and decoder prior to fusion. For example, the semantic gap between ($X^{0,0}$,$X^{1,3}$) is bridged using a dense convolution block with three convolution layers. In the graphical abstract, black indicates the original U-Net, green and blue show dense convolution blocks on the skip pathways, and red indicates deep supervision. Red, green, and blue components distinguish UNet++ from U-Net. (**b**) Detailed analysis of the first skip pathway of UNet++. (**c**) UNet++ can be pruned at inference time, if trained with deep supervision [29].

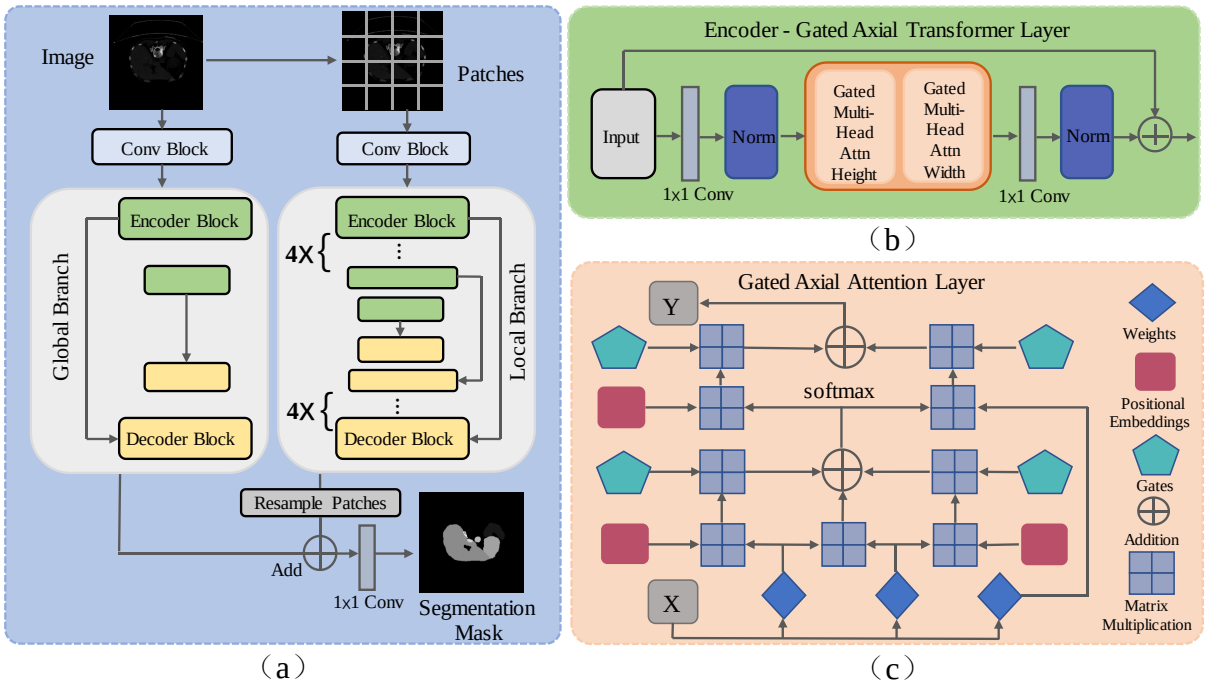

Figure 2. The Working Principle of MedT Architecture and Its Gated Axial Attention Mechanism. (**a**) The main architecture diagram of MedT which uses LoGo strategy for training. (**b**) The gated axial transformer layer which is used in MedT. (**c**) Gated Axial Attention layer which is the basic building block of both height and width gated multi-head attention blocks found in the gated axial transformer layer [38].

2.5. Cloud Computing in Medical Image Segmentation

As medical imaging technology continues to advance rapidly and the volume of large-scale medical imaging data grows, the role of cloud computing in medical imaging is expanding and evolving. Recently, the application of deep learning in medical imaging has garnered significant attention, facilitating the automatic analysis and diagnosis of medical imaging data. As a result, many researchers are investigating the integration of deep learning with cloud computing to efficiently process and manage large-scale medical imaging datasets.

Jimenez-del-Toro et al. [44] developed a cloud-based evaluation framework for assessing advanced anatomical structure segmentation methods. Chang [45] proposed a novel approach to brain segmentation that integrates medical education and research, exploring the advantages of cloud computing for segmentation both in terms of technology and user evaluation. Trägårdh et al. [26] and Egger et al. [46] created online scientific cloud platforms to foster collaboration between medical imaging and deep learning, particularly for tasks like organ segmentation. With the support of cloud computing, Shaukat et al. [47] improved brain tumor segmentation results using deep 3D U-Net.

2.6. Lightweight Segmentation Methods

As model size and complexity increase, so do computing and storage costs, which can limit the practical deployment of these models in resource-constrained environments. To overcome this challenge, researchers have focused on developing lightweight segmentation networks that enable efficient visual processing. MobileNets [48] introduced separable convolutions, combining depthwise (DW) and pointwise (PW) convolutions to extract feature maps. This approach significantly reduces the number of parameters and computational costs compared to traditional convolution operations. Grouped convolutions, initially used in the AlexNet [49] architecture to address memory limitations, also contribute to improving resource efficiency. These innovations have paved the way for more practical models that can function effectively in settings with limited computing and memory resources.

In the field of lightweight segmentation network research, MobileViT stands out as a significant advancement by combining the strengths of MobileNetV3 [50] and ViT [51] to create an efficient image segmentation method. MobileViT [52] is the first lightweight Transformer model designed for mobile devices, offering a novel approach by integrating Transformer and CNN architectures. Similarly, SegMarsViT [53] utilizes an encoder-decoder structure for segmentation tasks, with the MobileViT backbone network extracting both local and global spatial features in the encoder. However, ViT-based networks still encounter challenges, particularly in effectively propagating spatial and channel details and ensuring task-specific accuracy. Additionally, there remains significant room for

improvement in reducing the number of parameters and computational demands.

Existing research often seeks to enhance performance by incorporating complex modules, yet it frequently overlooks the constraints imposed by limited medical equipment resources in remote areas. Future research will likely shift focus towards developing medical image segmentation models that are low-parameter and computationally efficient.

3. The Overview of Brain Tumor Segmentation

3.1. Generative Model-Based Methods

Generative model-based methods focus on the appearance characteristics of both tumorous and healthy tissues, relying on domain-specific prior information, typically sourced from probabilistic image atlases. Menze et al. [54] enhanced a probabilistic atlas of healthy tissue priors with a latent atlas of lesions, and developed an estimation algorithm to extract tumor boundaries and the latent atlas from image data. Heinrich et al. [55] utilized discrete optimization and self-similarity within a discrete medical image registration framework for multimodal medical image segmentation.

3.2. Discriminative Model-Based Segmentation Methods

Discriminative model-based methods approach tumor segmentation as a classification problem, aiming to determine the properties of voxels [56]. With the rapid advancements in machine learning techniques, these methods have become the dominant approach in the field. Early methods in this category primarily relied on hand-crafted features, such as local histograms [57] and texture features [58], and used discriminative models like decision trees [59] and conditional random fields [60] for classification.

3.3. Cnn-Based Segmentation Models

In recent years, deep learning methods have shown great success in addressing various computer vision challenges, including medical image segmentation tasks, as demonstrated in Figure 3 from [61], and brain tumor segmentation [62, 63]. CNNs have significantly impacted the medical imaging field due to their ability to learn complex representations in a data-driven manner. Early approaches often used patch-based classification strategies, where CNNs predicted the class of the center voxel within a 2D or 3D patch [56, 64]. However, these patch-based methods struggle to capture correlations among neighboring patches over large regions. To overcome this limitation, end-to-end semantic segmentation models such as U-Net [28], attention U-Net [33], and U-Net++ [65] have become widely used for brain tumor segmentation. U-Net [28] employs a classical encoder-decoder architecture and leverages data augmentation for end-to-end training, which is especially beneficial when segmentation training samples are limited. The development of U-Net has greatly advanced medical image segmentation algorithms, particularly in brain tumor segmentation. Various U-Net variants, including UNet++ [65] and Res-UNet [66], have further improved its performance. Myronenko [67] introduced a segmentation network that incorporates a variational autoencoder branch to reconstruct the input image for better feature learning. Liu et al. [68] added a Variational Autoencoder (VAE) decoder to reconstruct input images and used image fusion as an additional regularization method to enhance feature learning. While CNN-based models perform well in 2D brain tumor image segmentation, MRI segmentation methods relying on slice-by-slice 2D networks overlook important 3D sequence and positional information. To address this, Isensee et al. [69] proposed an adaptive framework combining 2D U-Net, 3D U-Net, and U-Net Cascade, which automatically adjusts all hyperparameters without human intervention.

Recently, there has been a surge in the development of 3D network models designed to leverage 3D spatial information and extract high-dimensional feature representations from 3D MRI data. Unlike 2D networks, which can only process individual slices and must balance sparse inter-slice information with dense intra-slice details, 3D networks can capture spatial relationships along the depth dimension. This allows them to better understand the spatial structure of tumors and their interactions with surrounding tissues. When processing volumetric data, 3D models preserve continuous spatial information more effectively, minimizing the loss of important details. This is particularly critical for tasks like brain tumor segmentation. 3D fully convolutional networks (3D FCN) [70] are commonly used in brain volume image segmentation, with prominent models such as 3D U-Net [71] and nnU-Net [69]. The nnU-Net framework has become a strong baseline for both 2D and 3D medical image segmentation, and its robust performance has led to the development of several brain tumor segmentation models and their variants. For example, CANet, proposed by Liu et al. [72], enhances feature extraction by incorporating feature interaction maps, which interact with convolutional spaces to capture discriminative features in context. Zhu [73] proposed a Dual-Branch Ultrasound Image Segmentation Network (DBUNet) with two encoding branches for segmenting the original ultrasound image and the enhanced ultrasound image. It highlights the regions of interest and compensates

for the information loss during the enhancement process by realizing the interaction between two images.

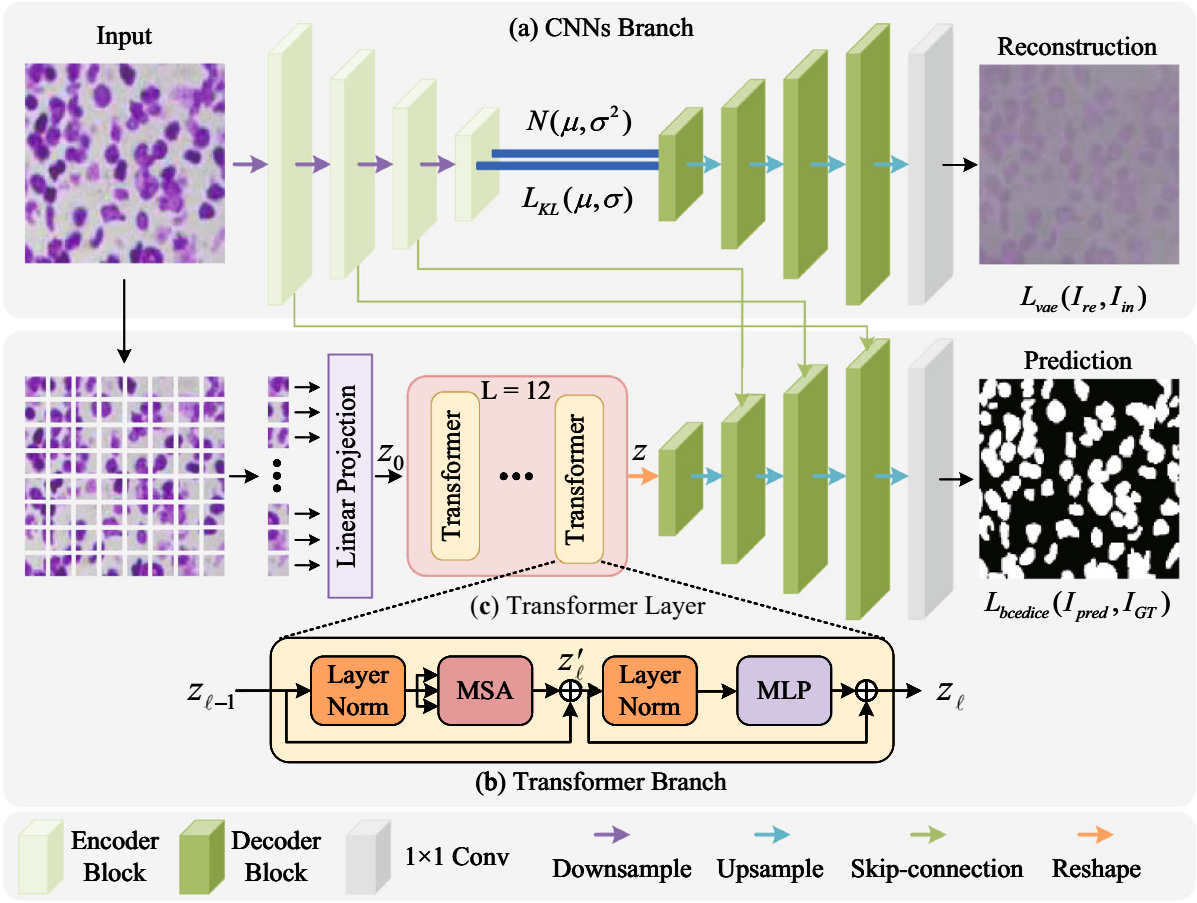

Figure 3. The dual encoding–decoding X-shaped network (X-Net) structure.(**a**) architecture of the CNNs branch, (**b**) architecture of the Transformer branch, (**c**) schematic of the Transformer layer [61]. Multi-head self-attention (MSA), and multilayer perceptron (MLP) blocks, layer normalization (LN) is applied before each block, and residual connection is applied after each block. MLP contains two fully connected layers with GELU (Gaussian Error Linear Unit) sub-linearity. L_{KL} is the standard VAE (variational auto-encoder) penalty item, used to estimate the KL (Kullback–Leibler divergence) dispersion between the normal distribution $N(\mu,\sigma^2)$ and the prior distribution $N(0,1)$. N is the total number of pixels in the image. μ and σ are the mean and standard deviation extracted by Gaussian distribution, respectively. z is the output by layer normalization. L is the number of layers. L_{vae} is an L2 loss used to match the VAE reconstructed image I_{re} with the input image I_{in}. $L_{bcedice}$ is the main loss function of the segmentation network used to match the segmentation prediction I_{pred} of Transformer branch and the ground truth (GT) I_{GT}.

3.4. Transformer-Based Segmentation Models

While CNN-based methods have achieved significant success in brain tumor segmentation, they are limited by their inability to capture global contextual information due to their restricted local receptive fields, which is crucial for semantic segmentation. In contrast, Transformers excel at modeling global interactions, a capability that CNNs struggle with due to the inherent constraints of convolution operations. As a result, Transformer-based methods have gained increasing attention in medical image segmentation, leading to the development of several notable models. Chen et al. [42] introduced TransUNet, a hybrid Transformer-CNN architecture, to explore the potential of Transformers in medical image segmentation. In this design, CNNs serve as feature extraction and transformation modules, while the Transformer handles global context encoding, as shown in Figure 4. MedT [38] introduced gated axial-attention specifically for medical image segmentation. Shi et al. [74] applied the Swin Transformer [35] concept and incorporated a simple yet effective Multi-layer Perceptron (MLP) decoder into a hierarchical SSformer for semantic segmentation. However, like standard ViT, Swin Transformer has limitations, particularly in terms of local context bias and the large computational resources it requires. Additionally, Transformer-based models often need to be pre-trained on large datasets such as ImageNet, which demands substantial computing power.

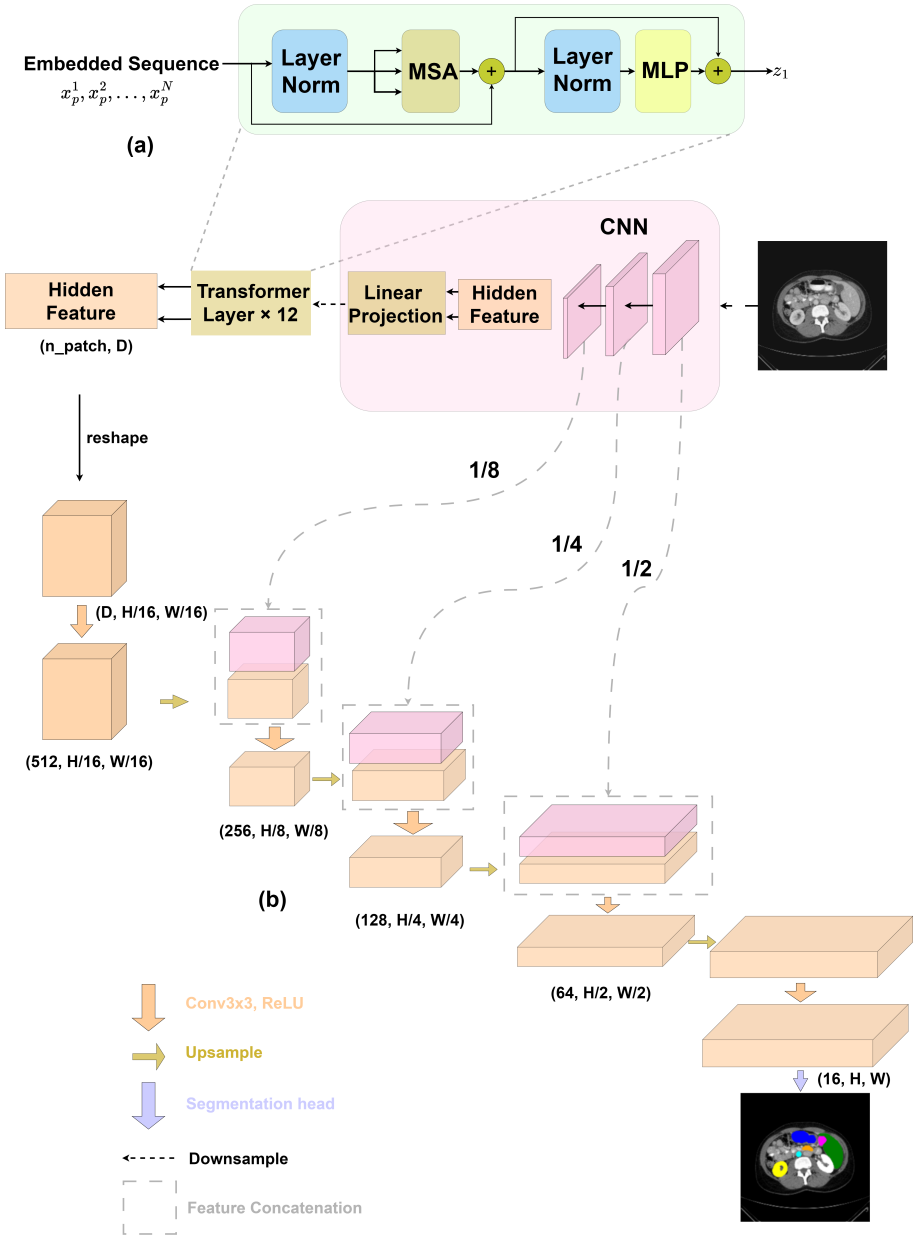

Figure 4. Overview of the framework. (**a**) schematic of the Transformer layer; (**b**) architecture of TransUNet [42].

However, directly segmenting the image into patches to serve as Transformer tokens results in the neglect of the local structure within the 3D volume. To effectively leverage 3D volumetric data for global interaction modeling between consecutive slices, Wang et al. [43] introduced TransBTS, the first approach to incorporate Transformer into a 3D CNN framework for 3D MRI brain tumor segmentation. Hatamizad et al. [75] proposed UNETR (UNEt TRansformers) , a ViT-based architecture for 3D medical image segmentation, which uses a pure Transformer encoder to capture the sequential representation of input data. The encoder is connected to a CNN-based decoder through skip connections, enabling the fusion of local and global information. However, standard ViT-based methods suffer from high computational complexity, especially for dense predictions like semantic segmentation, due to their fixed input size [36]. To address this, Swin Transformer [35] adopts a hierarchical structure that reduces computational complexity while effectively enhancing feature mapping. This modification significantly improves the performance of Transformer models in medical image segmentation tasks. Zhou et al. [76] introduced nnFormer, a 3D Transformer block-based model that interleaves convolution and self-attention operations, utilizing skip attention to concatenate encoder and decoder features. Swin UNETR [77] utilizes a Transformer-based encoder to learn multi-scale contextual representations and model long-range dependencies. Peiris et al. [78] developed VT-UNet, a lightweight UNet-shaped architecture that segments 3D medical images hierarchically by encoding both local and global features through a volumetric Transformer, as depicted in Figure 5. The core formulas governing the encoding and decoding processes are as follows:

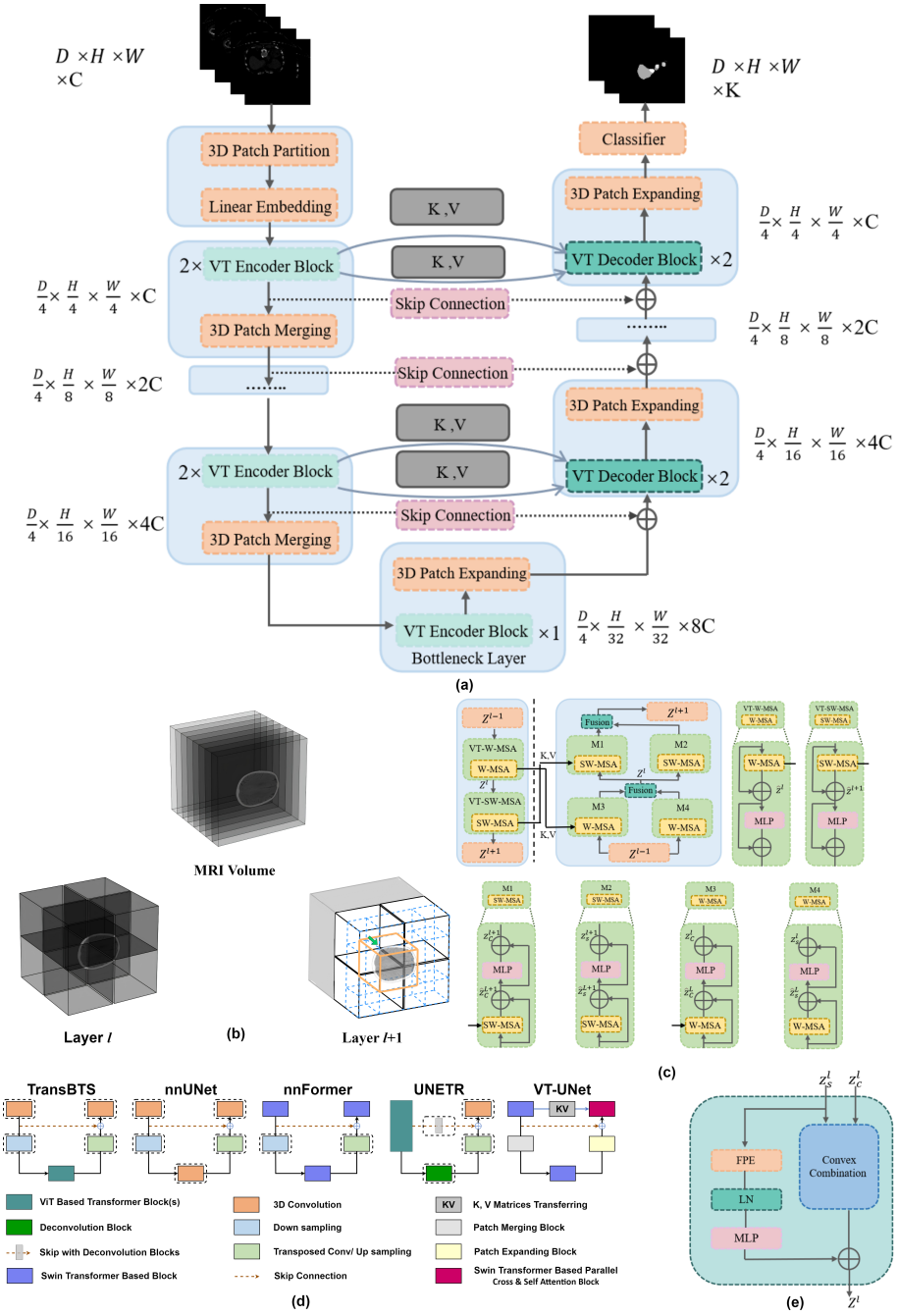

Figure 5. Architecture Overview of VT-UNet for 3D Medical Image Segmentation, Illustrating Volumetric Patch Partitioning, Encoder-Decoder Blocks, and Fusion Mechanism (**a**) Illustrates VT-UNet Architecture. Here, k denotes the number of classes. (**b**) shows visualization of Volumetric Shifted Windows. Consider an MRI volume of size $D \times H \times W$ with $D = H = W = 8$ for the sake of illustration. Further, let the window size for partitioning the volume be $P \times M \times M$ with $P = M = 4$. Here, layer l adopts the regular window partition in the first step of Volumetric Transformer(VT) block which results in $2 \times 2 \times 2 = 8$ windows. Inside layer $l + 1$, volumetric windows are shifted by $(\frac{P}{2}, \frac{M}{2}, \frac{M}{2})$ = (2, 2, 2) tokens. This results in $3 \times 3 \times 3 = 27$ windows. (**c**) shows VT Encoder-Decoder Structure. (**d**) Encoder-Decoder structural comparison with other SOTA methods. The proposed VT-UNet architecture has no convolution modules and is purely based on Transformer blocks. (**e**) Illustrates the structure of the Fusion Module [78].

Encoder Block:

$$\begin{aligned} \hat{z}^l &= \text{VT-W-MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1}, & \hat{z}^{l+1} &= \text{VT-SW-MSA}\left(\text{LN}\left(z^{l}\right)\right) + z^{l}, \\ z^l &= \text{MLP}\left(\text{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l, & z^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \end{aligned} \tag{3}$$

Decoder Block:

$$\text{SA}_r = \text{SA}(\mathbf{Q}_D, \mathbf{K}_D, \mathbf{V}_D), \quad \text{CA}_l = \text{SA}(\mathbf{Q}_D, \mathbf{K}_E, \mathbf{V}_E) \tag{4}$$

where $\hat{z}^l$ and $\hat{z}^{l+1}$ are intermediate features after VT-W-MSA (window-based self-attention) and VT-SW-MSA (shifted window-based self-attention), respectively, and z^l and z^{l+1} are final feature maps refined by MLP (multi-layer perceptron). $\mathrm{LN}(\cdot)$ denotes layer normalization applied to input features, VT-W-MSA and VT-SW-MSA capture local-global context within fixed and shifted 3D windows, and MLP enhances feature representation. These operations facilitate hierarchical feature encoding for efficient 3D segmentation. The encoder captures both local and global context through self-attention, while cross-attention integrates the encoder's features for refined segmentation, ultimately enhancing the accuracy of 3D medical image segmentation. The Decoder Block uses two primary mechanisms: Self-Attention (SA) and Cross-Attention (CA). In SA, the decoder uses its own queries (Q), keys (K), and values (V) to refine the feature map from the previous block. Specifically, SA_r employs the decoder's own queries, keys, and values. In CA_l, the decoder's queries interact with keys and values from the encoder, allowing the decoder to leverage both local and global information for accurate segmentation.

Despite its advantages, the Swin Transformer, like the standard ViT, suffers from a lack of locality inductive bias [79]. This limitation makes it challenging to apply the Swin Transformer to small datasets without pretraining, which can be problematic in medical image analysis tasks, such as brain tumor segmentation, where suitable pre-trained models may not always be available. To address this issue, a shifted patch tokenization strategy [79] was introduced into the Swin Transformer for brain tumor segmentation, enabling the model to be trained from scratch.

3.5. Multi-Path and Local Feature Fusion-Based Segmentation Methods

To achieve more accurate segmentation, most existing methods facilitate the interaction of global semantic features and local features by incorporating multi-path fusion learning or local information fusion modules. Chandrakar et al. [80] proposed a multipath CNN architecture for brain tumor segmentation and detection, enabling the fusion of local and global features. Zhao et al. [81] introduced a deformable multi-path ensemble (D-MEM) for automatic segmentation, combining both local and global features. While multipath fusion enhances feature learning, it is computationally expensive as it requires calculating all local and global paths. This makes it less feasible for 3D brain tumor segmentation due to high memory demands. In contrast, Wang et al. [82] proposed a Transformer-based MISSU (Medical Image Segmentation via Self-distilling TransUNet) model, which utilizes self-distillation and a local multi-scale fusion module to capture details from encoder skip connections while learning both global semantic information and local spatial features. Zhang et al. [83] introduced a multipath feature fusion module and a multichannel feature pyramid module to capture information from small targets. Zhou et al. [84] proposed a method for lossless feature computation in brain tumor segmentation, using 3D atrous convolutional layers and a coarse convolutional feature pyramid to combine background and lesion information. Although local information fusion modules use less computational memory, they may struggle to fully capture multi-scale details and edge features. The effective selection and fusion of detail features, such as edges and textures, are crucial for improving image information learning [85].

Although the CNN- and Transformer-based 3D medical image segmentation methods mentioned above have demonstrated impressive segmentation performance, they mainly focus on learning global spatial features from 3D volumetric data, often neglecting the detailed feature representations at various levels and resolutions. The proposed solution addresses this by not only learning global spatial features along the spatial axes and emphasizing inter-layer feature information, but also by focusing on extracting local features and edge features from multiple tumor regions within the volume. This multi-layer approach ensures that edge information is incorporated, which is crucial for accurately identifying and delineating tumor locations. By merging the extracted edge features with global spatial features, segmentation accuracy is enhanced, leading to improved overall segmentation performance.

3.6. Multimodal Segmentation Methods

The Brain Tumor Segmentation Challenge (BraTS Challenge), organized by the Medical Image Computing and Computer-Assisted Intervention Society (MICCAI), is a widely recognized public dataset used for brain tumor segmentation tasks. It has become a standard benchmark for evaluating the performance of brain tumor segmentation algorithms. Researchers and developers use the BraTS dataset to validate and compare the effectiveness of various algorithms, advancing the field of brain tumor segmentation. Over the years, numerous innovative tumor image segmentation network architectures have been proposed in the BraTS challenge. In 2016, Kamnitsas et al. [64] introduced an efficient fully connected multi-scale CNN framework called DeepMedic, which recombined high- and low-resolution paths to achieve better segmentation results. The following year, Wang [86] proposed a cascading architecture that decomposed the multi-class segmentation task into three binary segmentation problems, sequentially segmenting three tumor regions with inclusion relationships. In 2019, Isensee et al. [87] modified the widely used U-Net architecture, incorporating the dice loss function and deep supervision to address class imbalance issues. The

champion of the 2018 BraTS challenge [67] enhanced the U-Net framework by adding an image reconstruction branch, forming a VAE structure. This additional reconstruction branch provided extra guidance and regularization to the encoder, improving the clustering of the encoder's output features. The champion of the 2019 challenge [88] proposed a two-stage approach, where the output from the first stage, along with the original image, was used as input for the second stage, enabling a coarse-to-fine segmentation process.

To achieve more accurate segmentation results, the use of multimodal MRI data has become a key area of focus in brain tumor segmentation. However, most existing methods merely stack multimodal MRI scans into a multi-channel input without fully addressing the varying significance of each modality in relation to tumor segmentation. Pereira et al. [89] developed a convolutional network for automatic brain tumor segmentation using a four-channel format for multimodal images. Dolz et al. [90] extended dense connections to multimodal image segmentation using DenseNets, where each modality was treated as a separate branch, and dense connections were used to fuse features from different modalities. Liu et al. [91] introduced an attention-based modality selection feature fusion module to refine multimodal features, addressing the differences in relevance among modalities for the segmentation task. Zhang et al. [92] utilized FCN to extract features from various modalities and designed a modality-aware module for efficient information exchange across them. Mo et al. [93] classified the modalities into primary and auxiliary types, applying attention mechanisms for feature fusion. Although these approaches make valuable strides in leveraging multimodal MRI data, they primarily focus on the extraction and selection of deep semantic features, often overlooking features that hold specific importance for segmentation. The method proposed by Zhang et al. [94] integrates tumor prototypes and multi-expert networks across modalities, which not only focuses on deep semantic features but also emphasizes the localization and classification of tumor sub-regions, addressing the limitations of ignoring specific features in traditional methods. Furthermore, the deformation-aware and reconstruction-driven method proposed by Li et al. [95] improves segmentation performance by extracting deformation-aware features and using reconstruction, especially when some modality data is missing. Therefore, in addition to semantic features, it is crucial to focus on the extraction of edge information from relevant modalities such as FLAIR and T1ce, as this edge information is essential for improving segmentation quality. It helps accurately locate and delineate tumor boundaries. By merging these edge features with the semantic features, the aim is to use multimodal MRI data more effectively and enhance segmentation accuracy.

3.7. Dimension Processing in Segmentation Methods

The use of deep learning-based neural networks in medical brain tumor image segmentation primarily involves two approaches: two-dimensional (2D, slice) and three-dimensional (3D, voxel) processing. The 2D method begins by slicing 3D voxel data into 2D images, typically along the z-axis, and then feeds these slices into a CNN for learning and training. The resulting segmentation results from each 2D slice are subsequently reconstructed into a 3D representation. A 2D image, which is a projection of the MRI scan onto a plane, typically represents a transverse or longitudinal section. This projection provides essential information about the tumor's location, size, and morphology. Early classical segmentation networks such as FCN and UNet were initially applied using this method. UNet effectively combines deep and shallow information in medical image segmentation through skip connections that link encoder and decoder features. Later, Jegou [96] proposed the Fully Convolutional DenseNet (FC-DenseNet), which added dense connectivity blocks to the UNet structure, modifying the way features are connected during the upsampling and downsampling process. Havaei [56] introduced a fast segmentation method using a novel two-channel cascade structure. Gu [97] proposed the Context Encoder Network (CE-Net), which mitigates information loss during pooling and convolution by incorporating a context/extractor into the traditional encoding/decoding structure. The two-dimensional image boundary problem was also addressed, with methods focusing on boundary preservation [98] to improve the network's sensitivity and stability at the edges of segmentation targets. These techniques maximize the use of the entire 2D slice information, leading to improved performance in medical image segmentation.

The information provided by a 2D image is limited to the plane of the slice, capturing only the tumor's appearance in that specific plane. This may not offer a complete view of the tumor's spatial distribution and morphology. In contrast, 3D images are constructed by stacking multiple 2D slices, allowing for a more comprehensive representation of the tumor's three-dimensional shape and spatial distribution.

Compared to 2D segmentation, 3D segmentation offers several advantages in two key areas. First, in terms of data format, 3D data provides more directional information than 2D data, offering a more comprehensive representation of the spatial distribution and three-dimensional morphology of brain tumors [99]. Medical images are often represented as 3D data with multiple stacked slices, incorporating more information along the z-axis. Second, in terms of the model, 3D convolution processes data in all three dimensions (x, y, z), while 2D convolution

can only capture two dimensions (x, y) [99]. Third, 3D images offer a stereoscopic effect, enabling more intuitive visualization of tumor morphology, location, boundaries, and relationships. This makes tumor characteristics easier to interpret. 3D processing methods use CNN networks to perform convolutional operations on full 3D volumes or partial 3D blocks. A typical 3D segmentation network, like the V-Net proposed by F. Milletari [100], can be considered as a 3D U-Net with residual modules. In medical brain tumor image segmentation, deep learning-based networks generally employ either 2D (slice-based) or 3D (volume-based) processing techniques. The 2D approach involves slicing 3D voxel data into 2D images, which are then processed using CNNs with 2D convolution operations for training.

Finally, the segmentation results of each 2D slice are combined to form a 3D volume. Early segmentation network architectures, such as FCN and U-Net, were initially explored using this approach. The U-Net architecture improves medical image segmentation by using skip connections to link encoder and decoder features, allowing for the effective integration of both deep and shallow information. Later, Jegou [96] introduced Fully Convolutional DenseNets (FC-DenseNet), which enhanced U-Net by adding dense connection blocks, altering the connection scheme during upsampling and downsampling processes. Havaei [56] developed a fast segmentation method that employed a novel dual-channel cascade structure. Gu [97] proposed the Context Encoder Network (CE-Net), which addresses information loss during pooling and convolution by incorporating a context/extractor into the traditional encoder-decoder framework. Additionally, the challenge of boundary detection in 2D medical images was examined, with techniques such as boundary preservation [98] improving the sensitivity and stability of networks to edges in segmented targets.

3.8. Segmentation Frameworks with Deep Nuanced Reasoning and Swin-T

With the rapid advancements in computer technology and computer-aided diagnostic systems, medical image segmentation has become a prominent area of research. Deep learning, a key branch of machine learning, has significantly contributed to this progress. The increase in computational power and the availability of large-scale medical datasets have propelled deep learning to the forefront, making it an essential tool in the medical field for tasks such as image segmentation, feature extraction, and classification [101]. Since 2014, deep learning algorithms have been extensively explored for brain tumor segmentation in the context of the BraTS competition [102]. A growing number of studies have successfully employed neural network-based models for medical image segmentation [42, 61, 103]. These models include traditional architectures like CNN, FCN [27], and U-Net [28], as well as more recent innovations such as ViT [51] and the Swin-T network [35].

Vijay et al. [104] introduced SPP-U-Net, where traditional residual connections were replaced with a combination of spatial pyramid pooling (SPP) and attention blocks. Kamnitsas et al. [105] proposed the ensemble of multiple models and architectures (EMMA), which combined predictions from various 3D convolutional networks such as DeepMedic [106], FCN [27], and U-Net [28]. Isensee et al. [107] utilized nnU-Net [69], a self-configuring framework that automatically adapts U-Net to specific datasets. They showcased robust performance by making minimal modifications to the conventional 3D U-Net and incorporating optimizations tailored for the BraTS dataset. Luu et al. [108] proposed an enhancement to nnU-Net, which included using a larger network, replacing batch normalization with group normalization, and adding axial attention mechanisms in the decoder. While these methods often rely on the U-Net architecture, which consists of traditional encoder-decoder convolutional modules, they face challenges in capturing global contextual information.

The Diffusion Probability Model (DPM) has become a prominent topic in recent computer vision research. Wu et al. [109] introduced the MedSegDiff method for brain tumor segmentation using DPM. They proposed dynamic conditional encoding to establish adaptive conditions for each sampling step and introduced a feature frequency parser (FF Parser) to mitigate the negative effects of high-frequency noise components during the process. While their work demonstrated the versatility of DPM, the segmentation accuracy in brain tumor segmentation tasks still requires further improvement.

Wang et al. [43] introduced TransBTS, the first attempt to apply Transformers to multi-modal MRI brain tumor segmentation, yielding promising results. Lin et al. [110] developed a clinically knowledge-driven brain tumor segmentation model called CKD (Clinical Knowledge-Driven) TransBTS, as shown in Figure 6. Unlike previous approaches, this model did not directly link all modalities but instead grouped the input modalities based on MRI imaging principles. It reorganized the modalities and incorporated a dual-branch hybrid encoder with modality-correlated cross-attention (MCCA) blocks to extract features from multi-modal images. Zhu et al. [111] proposed a multi-task learning framework combining CNN and Swin-T, featuring a semantic segmentation branch and an edge detection branch. This framework aimed to leverage the strengths of different modalities, enhancing segmentation accuracy through complementary fusion and the segmentation of multi-modal features. However, these

approaches focus on complex network designs, leading to large model sizes, higher computational requirements, and longer training times.

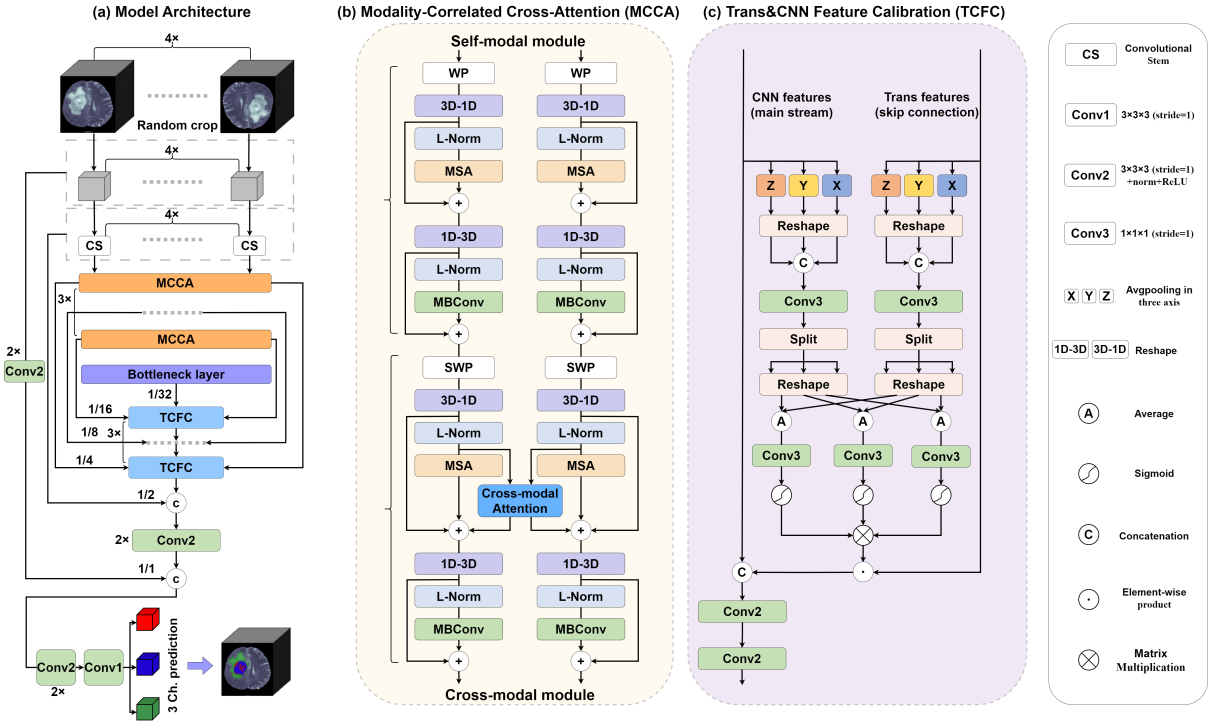

Figure 6. The architecture of CKD-TransBTS. (**a**) This model is a U-Net-like structure with a dual-branch hybrid encoder and a feature calibration decoder. According to the clinical knowledge of the MRI in brain tumor diagnosis, the input images are separated into two groups (T1 & T1Gd) and (T2 & T2FLAIR). Convolutional stem is introduced at the beginning. The encoder comprises several MCCA blocks ((**b**) Modality-Correlated Cross-Attention) which enables cross-modal interactions in a reasonable manner. The decoder consists of several TCFC blocks ((**c**) Trans&CNN Feature Calibration) to bridge the semantic gap between the features extracted by transformer and CNN. After several convolutional blocks, the model predicts the final brain tumor segmentation results. Note that, in the encoding (decoding) phase, the feature maps are downsampled (upsampled) by a convolutional (deconvolutional) layer at the end of each stage. In this figure, the downsample and upsample operations are omitted for simplification. The resolutions of the feature maps are specified at each stage by the scaling factors [110].

4. Performance Comparison

4.1. Brain Tumor Segmentation Based on the Fusion of Deep Semantics and Edge Information

4.1.1. Dataset and Implementation Details

The experiments [111] use training and testing datasets from the BraTS2018, BraTS2019, and BraTS2020 benchmarks [112–114]. BraTS, a prominent public dataset for multimodal brain tumor segmentation, is integral to the annual MICCAI brain tumor segmentation challenge and is widely used in related research. Each year's competition enhances the dataset by adding, removing, or replacing samples to expand its scope. Specifically, BraTS2018, 2019, and 2020 contain 285, 335, and 369 annotated brain tumor samples for model training, respectively. Each sample includes MRI scans from four modalities FLAIR (Fluid Attenuated Inversion Recovery), T1 (T1-weighted), T1ce (Contrast Enhanced T1-weighted), and T2 (T2-weighted), with annotations provided by domain experts. The labels include four categories: background, NCR/NET (Necrosis and Non-enhancing Tumor), ED (Edema),and ET (Enhancing Tumor). Evaluation is performed on three tumor regions: Whole Tumor (WT = NCR/NET + ED + ET), Tumor Core (TC = NCR/NET + ET), and Enhancing Tumor (ET). Performance is assessed using two widely adopted metrics in medical image segmentation: the Dice Score and the 95% Hausdorff Distance (HD).

During the preprocessing stage, each scan has a size of 240 × 240 × 155. The scans from all modalities are sliced, with each slice having a size of 240 × 240. For the semantic segmentation module, all four modalities are used as input. In the edge detection module, the input includes only the FLAIR and T1ce modalities. Additionally, z-score normalization, a commonly used technique, is applied to the raw data to address inconsistencies in image contrast across different modalities.

All programs were implemented using the PyTorch framework. The training process was carried out on four

Tesla P100 GPUs. The Adam optimizer [115] was used for the experiments, with a momentum value set to 0.9. The initial learning rate, weight decay, and batch size were set to 1×10^{-3}, 1×10^{-5}, and 16, respectively.

4.1.2. Comparison with Other Methods

To assess the effectiveness of the brain tumor segmentation method proposed by Zhu [111], several state-of-the-art segmentation techniques that have been evaluated on the BraTS2018-2020 benchmarks are used for comparison. These include 2D and 3D CNN-based methods [33, 65, 67, 71, 87, 88, 97, 116, 117], Transformer-based methods [42, 43], and methods focusing on multimodal feature fusion [118–120]. A brief overview of these methods is provided in Table 1. Since the source codes for many existing brain tumor segmentation methods are not publicly available, and to avoid biases introduced by model re-training, the evaluation results for these methods are directly taken from their respective publications, which is a standard practice in brain tumor segmentation research. The evaluation results for each method on the BraTS2018, BraTS2019, and BraTS2020 benchmarks are presented in Table 1, with the best-performing values highlighted in bold. The results are further visualized for comparison in Figures 7 and 8, which display the performance of different segmentation methods based on the Dice and HD metrics, respectively. The top-performing method in each case is marked with a star on the corresponding bar.

Table 1. Evaluation results of different brain tumor segmentation methods on the BraTS (2018–2020) datasets.

Datasets	Methods	WT		TC		ET		Average	
		Dice	HD	Dice	HD	Dice	HD	Dice	HD
BraTS2018	Myronenko [67]	90.40	4.483	85.90	8.278	81.40	3.805	85.90	5.500
	NoNewNet [87]	90.80	4.790	84.32	8.160	79.59	3.120	84.90	5.357
	U-Net++ [65]	88.96	5.327	84.65	8.535	79.49	4.285	84.36	6.049
	CENET [97]	89.53	5.271	84.31	8.493	79.95	4.379	84.60	6.193
	D. Zhang [118]	89.60	5.733	82.40	9.270	78.20	3.567	83.40	6.190
	TransUnet [42]	90.25	4.390	87.19	5.539	80.41	3.731	85.95	4.553
	Point-UNet [116]	90.55	-	87.09	-	80.76	-	86.13	6.010
	Z. Zhu [111]	**90.89**	**3.923**	**87.96**	**5.217**	**81.94**	**3.440**	**86.93**	**4.193**
BraTS2019	Attention Unet [33]	88.81	7.756	77.20	8.258	75.96	5.202	80.66	7.072
	U-Net++ [65]	89.67	6.345	87.13	5.521	80.25	3.313	85.68	5.060
	Z. Jiang [88]	90.94	4.263	86.47	5.439	80.21	3.146	85.87	4.283
	N3D [71]	91.60	6.547	88.80	6.219	83.00	3.543	87.80	5.436
	HNF-Net [117]	91.11	4.136	86.40	5.250	80.96	3.490	86.16	4.292
	T. Zhou [119]	89.70	6.700	77.50	9.300	70.60	7.400	79.27	7.800
	TransBTS [43]	90.00	5.644	81.94	6.049	78.93	3.736	83.62	5.143
	Z.Zhu [111]	**91.58**	**3.866**	**89.24**	**5.118**	**83.84**	**3.080**	**88.22**	**4.021**
BraTS2020	U-Net++ [65]	89.77	6.299	85.57	5.483	79.83	4.328	85.06	5.370
	Point-UNet [116]	89.67	-	82.97	-	76.43	-	83.02	8.260
	TransBTS [43]	90.09	4.964	81.73	9.769	78.73	17.947	83.52	10.893
	RFNet [120]	91.11	-	85.21	-	78.00	-	84.77	-
	Z. Zhu [111]	**91.03**	**4.719**	**88.22**	**5.985**	**84.61**	**3.051**	**87.95**	**4.585**

Based on the results presented in the above Tables and Figures, the method proposed by Zhu [111] demonstrates superior performance compared to other methods. Specifically, for the average Dice score, Zhu's method [111] achieves 86.71%, 88.22%, and 87.95% on the BraTS2018-2020 benchmarks, outperforming other reference methods by margins ranging from 0.58% to 7.81%, 0.42% to 8.95%, and 2.89% to 4.93%, respectively. When compared to TransBTS, which combines Transformer and U-Net for semantic segmentation, Zhu's method [111] consistently delivers better results, with particularly notable improvements in the tumor core region. Additionally, compared to the latest RFNet method, which incorporates multimodal feature fusion, Zhu's method [111] shows significant improvements in both the tumor core and enhancing tumor regions.

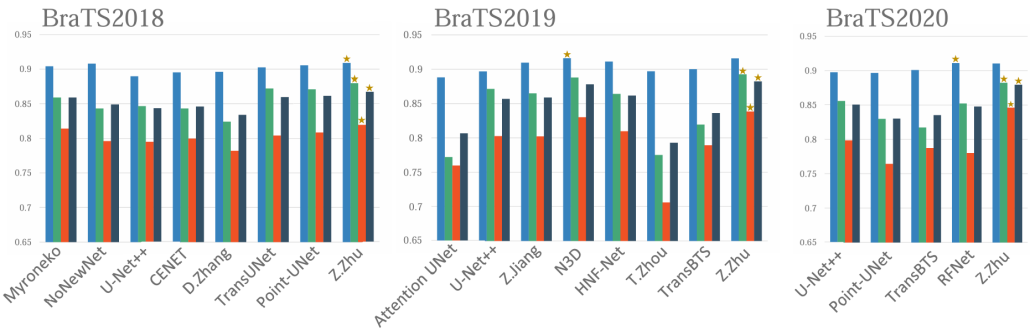

Figure 7. Performance comparison of different brain tumor segmentation methods on the metric Dice. The best-performed method in each case is marked by a star [111].

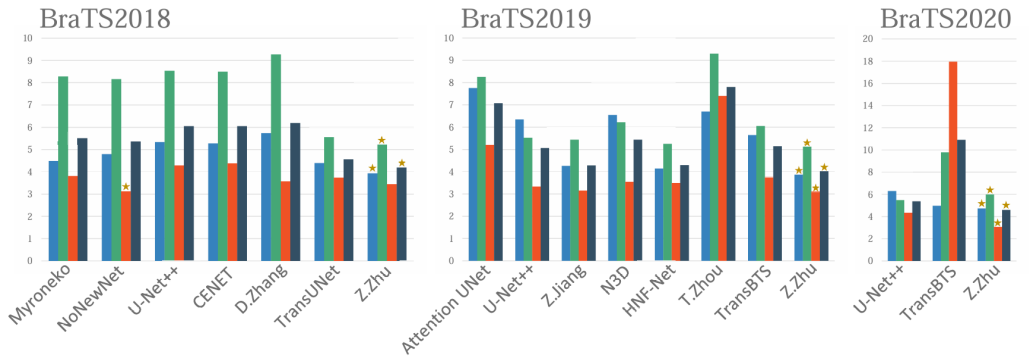

Figure 8. Performance comparison of different brain tumor segmentation methods on the metric HD. The best-performed method in each case is marked by a star [111].

Figure 9 presents a visual comparison of the brain tumor segmentation results produced by different methods. By referencing the ground truth, it is evident that the method proposed by Zhu [111] delivers more accurate segmentation results, particularly in delineating the tumor edges, when compared to other methods. This highlights the effectiveness of the edge features extracted for segmentation.

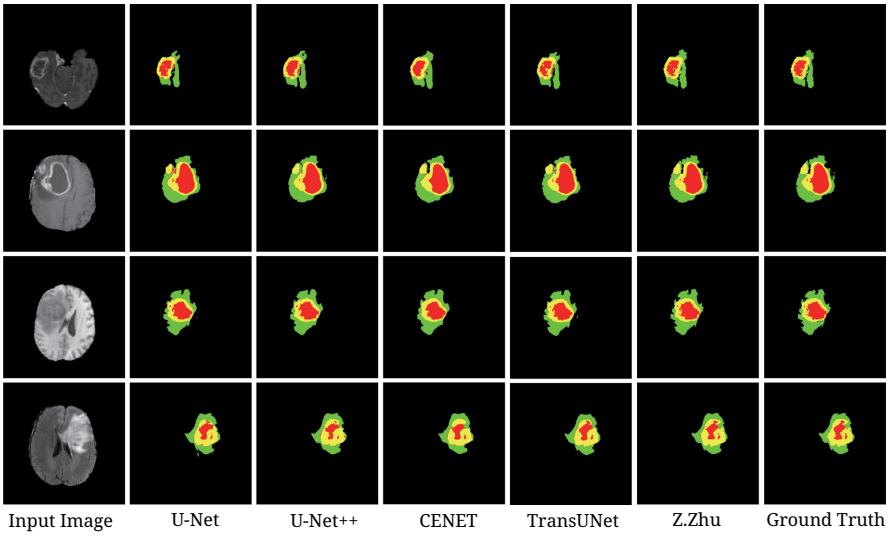

Figure 9. Visual effect comparison of brain tumor segmentation results obtained by different methods. The green, yellow and red indicate ED, ET and NCR/NET regions, respectively. (For interpretation of references to color in this figure legend, readers are referred to the online version of this article) [111].

Figure 10 provides an example comparing the performance of different segmentation methods in terms of tumor boundary accuracy. As noted earlier, the Hausdorff Distance (HD) metric is particularly sensitive to boundary shape, and the corresponding HD scores for the whole tumor are also presented. Among all the methods, Zhu's approach [111] achieves the best results, both in terms of HD scores and visual quality. These findings further emphasize that incorporating edge features enhances the accuracy of brain tumor segmentation.

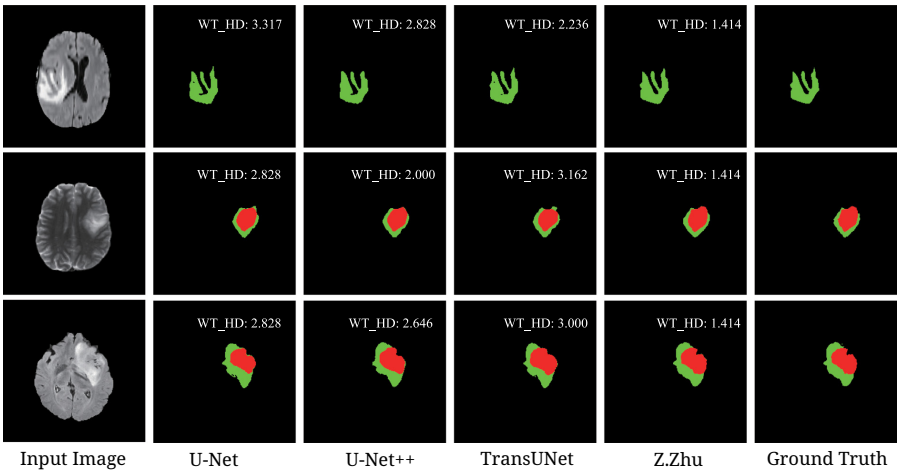

Figure 10. Performance comparison of different segmentation methods in terms of tumor boundary accuracy [111].

4.2. Lightweight Medical Image Segmentation Network

4.2.1. Dataset and Implementation Details

The experiments [121] were conducted on the ISIC2017 [122] and ISIC2018 [123] datasets. The ISIC (International Skin Imaging Collaboration) datasets are widely recognized and openly available in dermatological research. These datasets aim to support computer-aided dermatology diagnosis and research by offering a large collection of skin lesion images along with relevant clinical metadata.

Each image in the dataset has a size of 2166 × 3188 pixels. The ISIC 2017 dataset includes a training set, validation set, and test set with 2000, 150, and 600 dermoscopic images, respectively. Similarly, the ISIC 2018 dataset contains training, validation, and test sets with 2594, 100, and 1000 skin lesion images, respectively. For consistency, all images in this experiment were resized to 256 × 256 pixels.

To evaluate the performance of the segmentation network using consistent metrics, Intersection over Union (IoU), Dice score, and Segmentation Accuracy (SA) are employed. Additionally, the method proposed by Zhu [121] is compared with the baseline in terms of both the number of parameters and computational complexity, measured in floating-point operations (FLOPs).

The definitions of the IoU, Dice score, and SA are given as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{5}$$

$$\text{Dice} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}} \tag{6}$$

$$\text{SA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. These metrics were used to evaluate the performance of the segmentation network. Additionally, the average time (in milliseconds) required by the method proposed by Zhu [121] to segment an image, along with its frames per second (FPS), was calculated to assess its efficiency in comparison to other methods.

All experiments were implemented using the PyTorch framework and executed on a desktop equipped with a 24.00 GB NVIDIA GeForce GTX 3090, an Intel Core i7-8700MQ CPU @ 3.20 GHz, and 48.00 GB of RAM. The Adam optimizer with a momentum of 0.9 was used in the experiments. The initial learning rate, weight decay, and batch size were set to 1×10^{-3}, 1×10^{-5}, and 16, respectively.

4.2.2. Comparison with Other Methods

In the comparative experiments, the proposed Lightweight Medical Image Segmentation network with a multi-scale feature interaction guidance fusion framework - extra small (LMIS-xxs)[121] was compared with state-of-the-art models such as robust and lightweight deep learning real-time Segmentation Network for Multi-modality Medical Images (MISegNet) [124], Multi-Scale Contextual Attention Network (MSCA-Net)for skin lesion segmentation [125], and others [126, 127], highlighting its significant advantages across various aspects. To visually emphasize the lightweight nature of LMIS-xxs, bar charts are used for comparison with these models. As shown

in Figure 11, the chart compares the number of model parameters and FLOPs of LMIS-xxs with other methods. LMIS-xxs stands out with the lowest number of model parameters and FLOPs. Notably, the proposed lightweight LMIS-xxs model has only 0.524 M parameters and 0.197 G FLOPs.

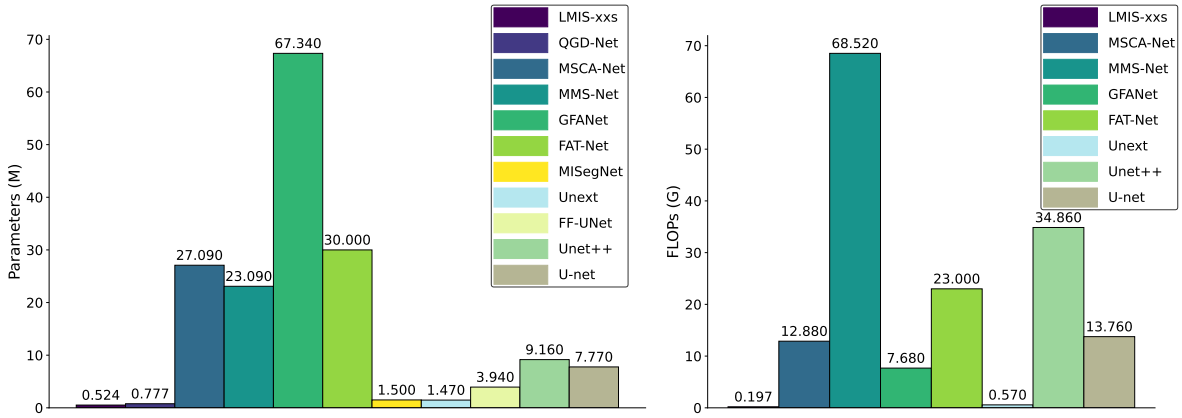

Figure 11. Histogram visualization comparison with other methods on parameters and FLOPs [121].

Table 2 presents a performance comparison of LMIS-xxs with other methods on the ISIC2017 and ISIC2018 datasets. The experimental results demonstrate that LMIS-xxs achieves state-of-the-art overall performance on the ISIC2017 dataset. Specifically, compared to larger U-Net models, LMIS-xxs not only delivers superior performance but also significantly reduces the number of parameters and FLOPs. When compared to other lightweight models, LMIS-xxs improved the IoU score by 9.75%, outperformed QGD-Net (Quaternion Group Dilated Neural Network), and showed a better balance between segmentation performance and model size. It also surpassed MSCA-Net, with reductions of 98.06% in parameters and 98.47% in computational effort. LMIS-xxs achieved optimal performance in terms of Segmentation Accuracy (SA). On the ISIC2018 dataset, LMIS-xxs matched the best-performing methods and outperformed most others. It outperformed UNeXt in terms of parameters, FLOPs, and segmentation performance. Notably, LMIS-xxs also performed better than MSCA-Net in terms of parameters, FLOPs, and Dice score, while MSCA-Net has a high computational complexity with 27.09 M parameters and 12.88 G FLOPs. Overall, LMIS-xxs offers competitive segmentation performance while providing significant advantages in terms of model parameters and computational complexity.

Table 2. Comparison of methods on ISIC2017 and ISIC2018 datasets [121].

Datasets	Methods	Parameters (M)	FLOPs (G)	IoU	Dice	SA
ISIC2017	U-Net [28]	31.13	55.840	76.18	84.92	91.64
	FAT-Net [126]	30.000	23.000	76.53	85.00	93.26
	MISegNet [124]	1.500	-	-	86.45	-
	GFANet [128]	23.090	7.680	77.75	85.74	93.97
	MMS-Net [129]	67.340	68.520	77.90	87.60	95.40
	MSCA-Net [125]	27.090	12.880	79.26	87.31	94.41
	QGD-Net [130]	0.777	-	72.23	-	93.01
	EIU-Net [131]	14.160	18.920	77.10	85.50	93.70
	LMIS-xxs [121]	0.524	0.197	82.33	89.62	95.72
ISIC2018	U-Net [28]	31.13	55.840	74.55	84.03	93.04
	FF-UNet [132]	3.940	-	80.20	88.70	96.40
	FAT-Net [126]	30.000	23.000	82.02	89.03	95.78
	UNeXt [133]	1.470	0.570	82.78	90.41	-
	MSCA-Net [125]	27.090	12.880	84.18	90.52	96.41
	GFANet [128]	23.090	7.680	83.66	90.13	96.29
	EIU-Net [131]	14.160	18.920	83.60	90.20	96.70
	SMTF [134]	3.100	2.190	81.10	88.75	95.72
	TransGuider [135]	16.130	-	82.68	89.48	-
	Zhu [127]	4.280	38.580	83.26	90.72	-
	LMIS-xxs [121]	0.524	0.197	83.23	90.85	96.33

To further demonstrate the reliability and effectiveness of LMIS-xxs, the publicly available source codes of several methods were used for comparison, including U-Net [28], MSCA-Net [125], GFANet [128], U-Net++ [29], and UNeXt [133]. These models were trained under the same conditions on the ISIC2017 dataset and tested on the test set. As shown in Figure 12, the curves display the loss and IoU values obtained at each epoch during training, as well as the val_loss and val_IoU derived from the validation set. Comparing the results with these five advanced and classical methods, LMIS-xxs demonstrates a faster loss reduction during training, achieving lower training loss and higher average IoU values after convergence. Similarly, on the validation set, LMIS-xxs outperforms the other methods with lower loss and higher average IoU, showing more stable convergence. These results further highlight the advantages of the proposed LMIS-xxs network.

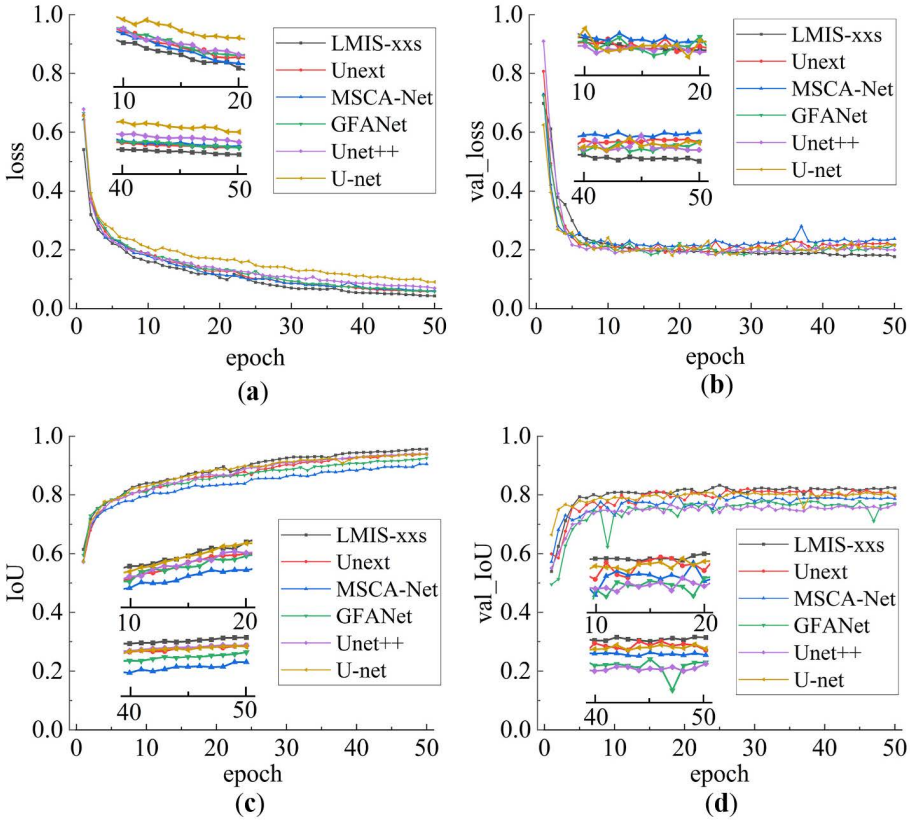

Figure 12. Visual performance comparison of LMIS-xxs and other methods in terms of (**a**) loss, (**b**) val_loss, (**c**) IoU and (**d**) val_IoU results. Loss and IoU of each epoch were obtained during the training process. Val_loss and val_IoU were obtained based on the verification dataset after the end of an epoch. These four indicators can be used to judge the learning status of the model during the training process [121].

Next, a visual comparison experiment was conducted on the ISIC2017 test set using these five methods, and box plots were used for statistical analysis. Figure 13 presents the visual experimental results comparing the LMIS-xxs network with these five methods. By comparing the results to the ground truth, it is clear that LMIS-xxs achieves more accurate segmentation than the other methods, effectively capturing dermatological regions of varying sizes. This demonstrates the superior segmentation performance of the proposed LMIS-xxs network.

During the model testing phase, each image in the test set was evaluated using the IoU and Dice metrics. Figure 14 shows box plot visualizations of these metrics for each image in the test set, comparing the segmentation performance of the LMIS-xxs network with five other methods. It is evident that LMIS-xxs outperforms the other methods in terms of segmentation performance, as indicated by the average values in the IoU and Dice box plots. Segmentation tasks often involve images with poor segmentation results, which are represented as outliers in the box plots. LMIS-xxs exhibits fewer such outliers and demonstrates a smaller distance from the minimum values, indicating fewer instances of poor segmentation results. This highlights the generalization and superiority of LMIS-xxs in skin disease segmentation compared to other methods. Finally, examining the range between the minimum and maximum values in the box plots, it is clear that LMIS-xxs has a smaller range, with most IoU and Dice results concentrated within it, reflecting its higher robustness.

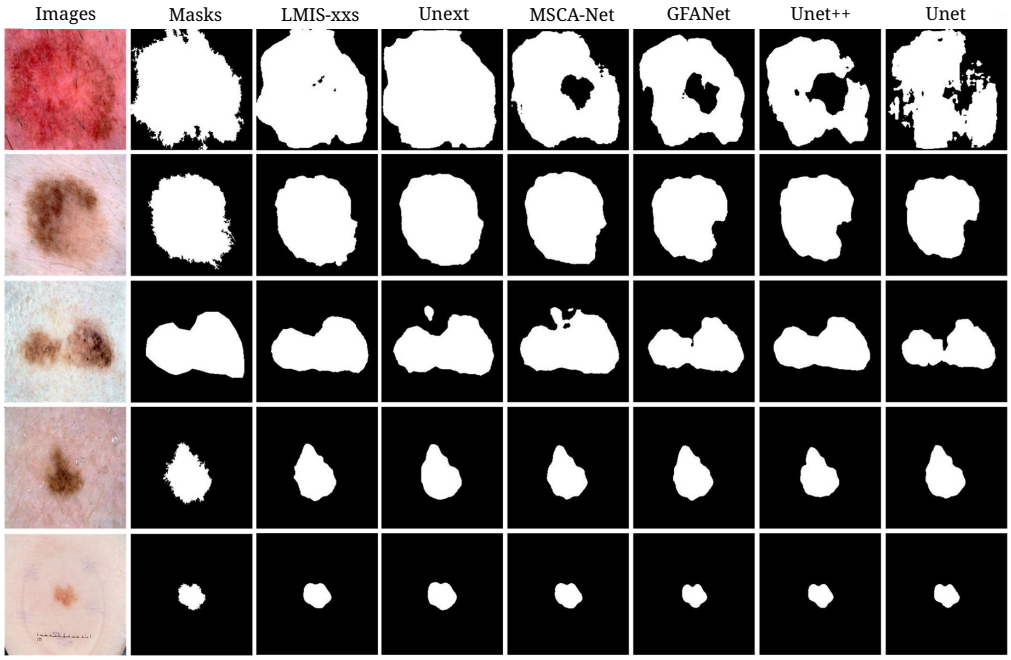

Figure 13. Visual effect comparison of dermatology segmentation results obtained by different methods. It shows selected skin disease segmentation cases in different size [121].

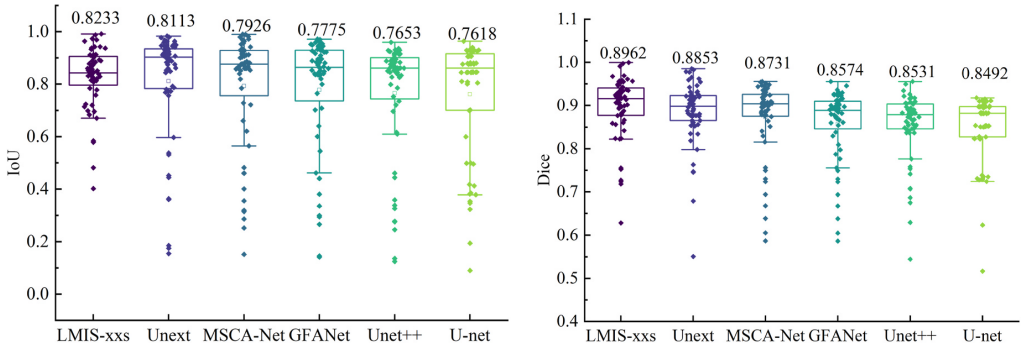

Figure 14. Box plots visualize IoU and Dice metrics for LMIS-xxs and five other methods on the test set. Each point in the box plot corresponds to the IoU and Dice evaluation for each image. Numerical values in the plot represent average IoU and Dice scores. The short horizontal line at the top of each box plot represents the upper bound (maximum value), while the line at the bottom represents the lower bound (minimum value). The size of the interval between the minimum and maximum values reflects the segmentation performance of the method. Data points below the corresponding minimum values represent outliers, indicating images with relatively poor segmentation results [121].

In summary, the experimental results highlight the lightweight nature of the LMIS-xxs segmentation model, which competes effectively with most other methods. On the ISIC dataset, LMIS-xxs demonstrates strong robustness in skin disease segmentation tasks, showcasing its advantages over other lightweight networks and making significant contributions to the overall experimental outcomes.

4.2.3. Extension to Other Medical Image Segmentation

To evaluate the adaptability of LMIS-xxs to various biomedical segmentation tasks, two additional biomedical image datasets were selected for testing: colorectal cancer lesion segmentation and cell nucleus segmentation. For colorectal cancer lesion segmentation, the CVC-ClinicDB dataset [136] and the Kvasir-SEG dataset [137] were used. The same training set as the polyp segmentation method [138] was utilized, consisting of 900 samples from Kvasir and 550 samples from CVC-ClinicDB for training, with the remaining images used for testing. The LMIS-L version of the proposed model was selected for this experiment. For cell nucleus segmentation, the 2018 DSB dataset [139], which includes 670 annotated images, was used. This dataset was randomly split into training and validation sets in an 8:2 ratio. For this experiment, the LMIS-xxs version of the model was chosen. All images were resized to 256 × 256 pixels. Table 3 presents the objective evaluation results of different methods on the

CVC-ClinicDB and 2018 DSB datasets, respectively. Figures 15 and 16 show the visual results of the proposed method for colorectal cancer lesion segmentation and cell nucleus segmentation.

Table 3. Comparison of methods on 2018 DSB, CVC-ClinicDB, and Kvasir-SEG datasets [121].

Datasets	Methods	Parameters (M)	FLOPs (G)	IoU	Dice
2018 DSB	U-Net [28]	31.13	55.840	82.57	88.93
	UNeXt [133]	1.470	0.570	82.15	88.13
	ConvUNeXt [140]	3.510	7.250	83.64	89.09
	TransAttUnet-C [141]	25.970	88.570	84.36	90.04
	Zhu [127]	4.280	38.580	83.26	90.72
	LMIS-xxs [121]	0.524	0.197	84.06	90.68
CVC-ClinicDB	U-Net [28]	31.13	55.840	75.50	82.30
	PraNet [138]	32.55	13.11	84.90	89.90
	MSNet [142]	29.74	16.97	87.90	92.10
	Polyp-PVT [143]	44.98	16.95	88.90	93.70
	MEGANet [144]	29.27	46.82	88.50	93.00
	LMIS-L [121]	3.552	1.126	88.34	92.28
Kvasir-SEG	U-net [28]	31.13	55.840	74.60	81.80
	PraNet [138]	32.55	13.11	84.00	89.80
	MSNet [142]	29.74	16.97	86.20	90.70
	Polyp-PVT [143]	44.98	16.95	86.40	91.70
	MEGANet [144]	29.27	46.82	85.90	91.10
	LMIS-L [121]	3.552	1.126	85.73	90.85

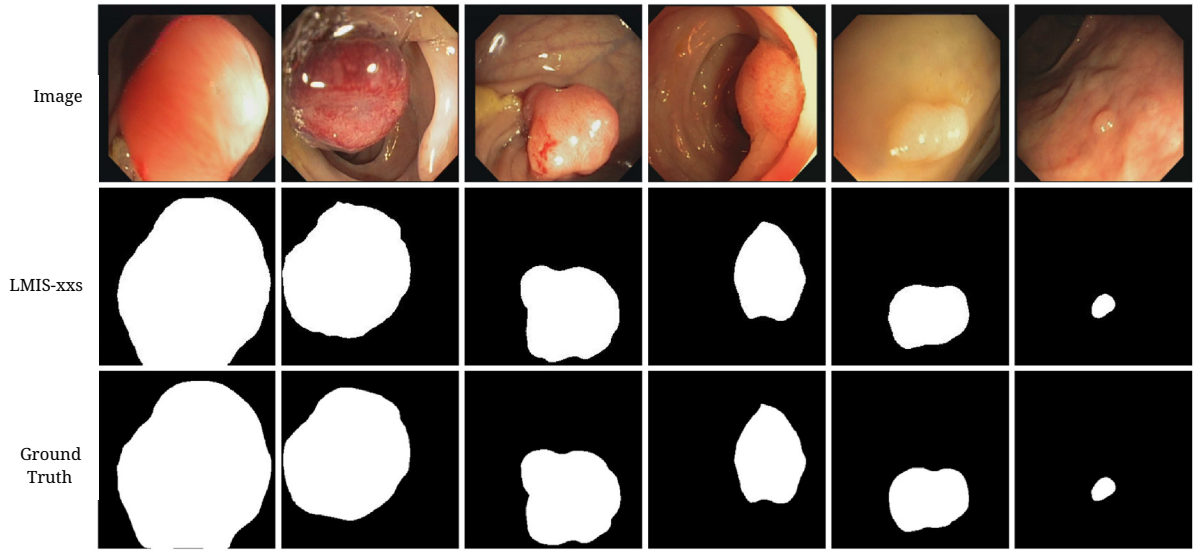

Figure 15. Visual effect comparison of colorectal cancer lesion segmentation results [121].

The results presented in Table 3 indicate that LMIS-xxs and LMIS-L outperform other methods in terms of competitiveness. Specifically, LMIS-L achieved average IoU and Dice scores of 88.34% and 92.28%, respectively, on the CVC-ClinicDB dataset. On the Kvasir-SEG dataset, the average IoU and Dice were 85.73% and 90.85%, respectively. When compared to other reference methods, LMIS-L outperformed PraNet [138] and MSNet [142], both of which were published in MICCAI, though it slightly lagged behind the MEGANet [144] method. However, LMIS-L demonstrated significant advantages in terms of model parameters and computational complexity, requiring 7.246 M to 44.426 M fewer parameters and reducing FLOPs from 11.984 G to 45.694 G compared to the other reference methods. On the 2018 DSB dataset, LMIS-xxs achieved average IoU and Dice scores of 84.06% and 90.68%, respectively. Compared to other methods, LMIS-xxs showed advantages in terms of model parameters and computational complexity. In terms of IoU, LMIS-xxs outperformed Zhu's method [127] published in *Pattern Recognition* by 0.8%, and in terms of Dice, it was 0.04–2.25% higher than the other methods. Figures 15 and 16

demonstrate that both LMIS-xxs and LMIS-L provide accurate segmentation results for colorectal cancer lesion segmentation and cell nucleus segmentation. In summary, LMIS-xxs and LMIS-L delivered highly satisfactory results in these two segmentation tasks, showing strong adaptability to other biomedical image segmentation challenges.

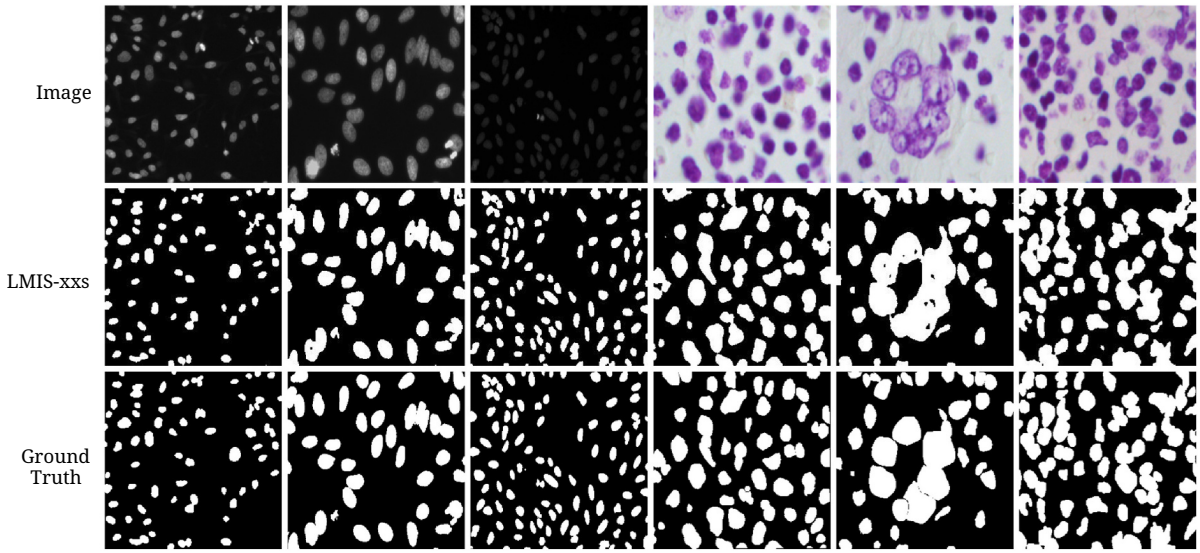

Figure 16. Visual effect comparison of cell nucleus segmentation results [121].

5. Conclusion

In conclusion, medical image segmentation continues to be a vital area of research with profound implications for clinical practice, such as diagnosis, treatment planning, and disease monitoring. Significant advancements have been made in segmentation techniques, evolving from traditional image processing methods to advanced deep learning approaches. While early segmentation tasks were based on classical methods like thresholding, region growing, and active contours, the advent of machine learning and deep learning has transformed the field, providing highly accurate, automated, and scalable solutions.

Deep learning, especially CNNs, has demonstrated remarkable potential in overcoming the challenges of medical image segmentation, including anatomical variability, image noise, and complex structures. However, despite these advancements, several challenges persist, such as limited data availability, class imbalance, model generalizability, and computational efficiency. Current models often rely on large, annotated datasets, which are costly and time-consuming to obtain, and may struggle to generalize across diverse populations or adapt to new imaging modalities.

Future research in medical image segmentation should aim to tackle these challenges through innovations like semi-supervised and unsupervised learning, enhanced data augmentation techniques, multi-modal segmentation, and transfer learning. Moreover, integrating explainability and interpretability into deep learning models is crucial for their clinical adoption, as healthcare professionals must be able to trust the results produced by these automated systems. With continuous advancements in artificial intelligence (AI), imaging technologies, and computational methods, the future of medical image segmentation looks promising, offering the potential to improve clinical outcomes and enable more personalized healthcare. The efficiency and effectiveness of medical image segmentation will be the future trend. The compatibility of lightweight network models and hardware devices will be focused on. Training results will be shared to reduce duplication of effort. AI tools such as ChatGPT, Gemini, and DeepSeek will be integrated into future medical image segemenation applications.

This survey provides a thorough overview of the current landscape of medical image segmentation, highlighting both its achievements and the ongoing challenges. As the field advances, interdisciplinary collaboration among computer scientists, medical professionals, and researchers will be crucial for developing robust, efficient, and clinically applicable segmentation models.

Author Contributions

G.Q., Z.Z., K.L. and H.X.: conceptualization, methodology, software; G.Q., K.L. and H.X.: data curation, writing—original draft preparation; K.L. and H.X.: visualization, investigation; Z.Z.: supervision; K.L. and H.X.: software, validation; G.Q., Z.Z., K.L. and H.X.: writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Patrascanu, O.S.; Tutunaru, D.; Musat, C.L.; et al. Future Horizons: The Potential Role of Artificial Intelligence in Cardiology. *J. Pers. Med.* **2024**, *14*, 656.
2. Xu, Y.; Yu, K.; Qi, G.; et al. Brain tumour segmentation framework with deep nuanced reasoning and Swin-T. *IET Image Process.* **2024**, *18*, 1550–1564.
3. Wahid, F.; Ma, Y.; Khan, D.; et al. Biomedical Image Segmentation: A Systematic Literature Review of Deep Learning Based Object Detection Methods. *arXiv* **2024**, arXiv:2408.03393.
4. Krikid, F.; Rositi, H.; Vacavant, A. State-of-the-Art Deep Learning Methods for Microscopic Image Segmentation: Applications to Cells, Nuclei, and Tissues. *J. Imaging* **2024**, *10*, 311.
5. Zhu, Z.; Yin, H.; Chai, Y.; et al. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inf. Sci.* **2018**, *432*, 516–529.
6. Zhu, Z.; Zheng, M.; Qi, G.; et al. A Phase Congruency and Local Laplacian Energy Based Multi-Modality Medical Image Fusion Method in NSCT Domain. *IEEE Access* **2019**, *7*, 20811–20824.
7. Wang, K.; Zheng, M.; Wei, H.; et al. Multi-Modality Medical Image Fusion Using Convolutional Neural Network and Contrast Pyramid. *Sensors* **2020**, *20*, 2169.
8. Gudigar, A.; Kadri, N.A.; Raghavendra, U.; et al. Automatic identification of hypertension and assessment of its secondary effects using artificial intelligence: A systematic review (2013–2023). *Comput. Biol. Med.* **2024**, *172*, 108207.
9. Rahmim, A.; Bradshaw, T.J.; Davidzon, G.; et al. Nuclear Medicine Artificial Intelligence in Action: The Bethesda Report (AI Summit 2024). *arXiv* **2024**, arXiv:2406.01044.
10. Hussain, S.I.; Toscano, E. An extensive investigation into the use of machine learning tools and deep neural networks for the recognition of skin cancer: Challenges, future directions, and a comprehensive review. *Symmetry* **2024**, *16*, 366.
11. Khandakar, S.; Al Mamun, M.A.; Islam, M.M.; et al. Unveiling Early Detection And Prevention Of Cancer: Machine Learning And Deep Learning Approaches. *Educ. Adm. Theory Pract.* **2024**, *30*, 14614–14628.
12. Sobur, A.; Rana, I.C. Advancing Cancer Classification with Hybrid Deep Learning: Image Analysis for Lung and Colon Cancer Detection. *SSRN* **2024**.
13. Zhu, Z.; Sun, M.; Qi, G.; et al. Sparse Dynamic Volume TransUNet with multi-level edge fusion for brain tumor segmentation. *Comput. Biol. Med.* **2024**, *172*, 108284.
14. Kong, Y.; Dun, Y.; Meng, J.; et al. A novel classification method of medical image segmentation algorithm. In Proceedings of the Medical Imaging and Computer-Aided Diagnosis: Proceeding of 2020 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2020), Oxford, UK, 20–21 January 2020; pp. 107–115.
15. Lin, Z.; Wang, Z.; Zhang, Y. Optimal evolution algorithm for image thresholding. *J. Comput. -Aided Des. Comput. Graph.* **2010**, *22*, 1201–1206.
16. Jianfeng, L. Adaptive region growing algorithm in medical images segmentation. *J. Comput. Aided Des. Comput. Graph.* **2005**, *17*, 2168.
17. Cai, H.; Xu, X.; Lu, J.; et al. Repulsive force based snake model to segment and track neuronal axons in 3D microscopy image stacks. *NeuroImage* **2006**, *32*, 1608–1620.
18. Sengur, A.; Guo, Y. Color texture image segmentation based on neutrosophic set and wavelet transformation. *Comput. Vis. Image Underst.* **2011**, *115*, 1134–1144.
19. Patil, D.D.; Deore, S.G. Medical image segmentation: A review. *Int. J. Comput. Sci. Mob. Comput.* **2013**, *2*, 22–27.
20. Guerrout, E.H.; Mahiou, R.; Ait-Aoudia, S. Medical image segmentation on a cluster of PCs using Markov random fields. *Int. J. New Comput. Archit. Their Appl.* **2013**, *3*, 35–44.

21. Cui, W.; Wang, Y.; Lei, T.; et al. Local region statistics-based active contour model for medical image segmentation. In Proceedings of the 2013 Seventh International Conference on Image and Graphics, Qingdao, China, 26–28 July 2013; pp. 205–210.
22. Li, B.N.; Chui, C.K.; Chang, S.; et al. Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. *Comput. Biol. Med.* **2011**, *41*, 1–10.
23. Saha, P.K.; Udupa, J.K.; Odhner, D. Scale-based fuzzy connected image segmentation: Theory, algorithms, and validation. *Comput. Vis. Image Underst.* **2000**, *77*, 145–174.
24. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep fisher networks for large-scale image classification. *Adv. Neural Inf. Process. Syst.* **2013**, *2013*, 26.
25. Wang, Z.; Li, H.; Ouyang, W.; et al. Learnable histogram: Statistical context features for deep neural networks. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
26. Zhu, L.; Ji, D.; Zhu, S.; et al. Learning statistical texture for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12537–12546.
27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
28. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015.
29. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; et al. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support 2018, Granada, Spain, 20 September 2018.
30. Xiao, X.; Lian, S.; Luo, Z.; et al. Weighted res-unet for high-quality retina vessel segmentation. In Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 19–21 October 2018; pp. 327–331.
31. Li, X.; Chen, H.; Qi, X.; et al. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674.
32. Jin, Q.; Meng, Z.; Sun, C.; et al. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. *Front. Bioeng. Biotechnol.* **2020**, *8*, 605132.
33. Oktay, O.; Schlemper, J.; Folgoc, L.L.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
34. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; et al. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, 17–21 October 2016.
35. Liu, Z.; Lin, Y.; Cao, Y.; et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
36. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
37. Li, H.; Su, D.; Cai, Q.; et al. BSAFusion: A Bidirectional Stepwise Feature Alignment Network for Unaligned Medical Image Fusion. *arXiv* **2024**, arXiv:2412.08050.
38. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; et al. Medical transformer: Gated axial-attention for medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021.
39. Guo, J.; Han, K.; Wu, H.; et al. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185.
40. Dai, Z.; Liu, H.; Le, Q.V.; et al. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.
41. Zheng, S.; Lu, J.; Zhao, H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 6881–6890.
42. Chen, J.; Lu, Y.; Yu, Q.; et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
43. Wang, W.; Chen, C.; Meng, D.; et al. Transbts: Multimodal brain tumor segmentation using transformer. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; pp. 109–119.
44. Jimenez-del Toro, O.; Müller, H.; Krenn, M.; et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. *IEEE Trans. Med. Imaging* **2016**, *35*, 2459–2475.

45. Chang, V. Cloud Computing for brain segmentation technology. In Proceedings of the 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, UK, 2–5 December 2013; Volume 1, pp. 499–504.
46. Egger, J.; Wild, D.; Weber, M.; et al. Studierfenster: An open science cloud-based medical imaging analysis platform. *J. Digit. Imaging* **2022**, *35*, 340–355.
47. Shaukat, Z.; Farooq, Q.u.A.; Tu, S.; et al. A state-of-the-art technique to perform cloud-based semantic segmentation using deep learning 3D U-Net architecture. *BMC Bioinform.* **2022**, *23*, 251.
48. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
49. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *2012*, 25.
50. Koonce, B.; Koonce, B.E. *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Springer: Berlin, Germany, 2021.
51. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4–9 December 2017; pp. 6000–6010.
52. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
53. Dai, Y.; Zheng, T.; Xue, C.; et al. SegMarsViT: Lightweight mars terrain segmentation network for autonomous driving in planetary exploration. *Remote Sens.* **2022**, *14*, 6297.
54. Menze, B.H.; Van Leemput, K.; Lashkari, D.; et al. A generative model for brain tumor segmentation in multi-modal images. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010: 13th International Conference, Beijing, China, 20–24 September 2010.
55. Heinrich, M.P.; Maier, O.; Handels, H. Multi-modal Multi-Atlas Segmentation using Discrete Optimisation and Self-Similarities. *Visc. Chall.* **2015**, *1390*, 27.
56. Havaei, M.; Davy, A.; Warde-Farley, D.; et al. Brain tumor segmentation with deep neural networks. *Med. Image Anal.* **2017**, *35*, 18–31.
57. Goetz, M.; Weber, C.; Bloecher, J.; et al. Extremely randomized trees based brain tumor segmentation. *Proceeding BRATS Chall. MICCAI* **2014**, *14*, 24.
58. Subbanna, N.K.; Precup, D.; Collins, D.L.; et al. Hierarchical probabilistic Gabor and MRF segmentation of brain tumours in MRI volumes. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, 22–26 September 2013.
59. Zikic, D.; Glocker, B.; Konukoglu, E.; et al. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, 1–5 October 2012.
60. Wu, W.; Chen, A.Y.; Zhao, L.; et al. Brain tumor detection and segmentation in a CRF (conditional random fields) framework with pixel-pairwise affinity and superpixel-level features. *Int. J. Comput. Assist. Radiol. Surg.* **2014**, *9*, 241–253.
61. Li, Y.; Wang, Z.; Yin, L.; et al. X-net: A dual encoding–decoding method in medical image segmentation. *Vis. Comput.* **2023**, *39*, 2223–2233.
62. Zhou, X.; Chen, Y.; Wu, Z.; et al. Boosted local dimensional mutation and all-dimensional neighborhood slime mould algorithm for feature selection. *Neurocomputing* **2023**, *551*, 126467.
63. Shi, B.; Chen, J.; Chen, H.; et al. Prediction of recurrent spontaneous abortion using evolutionary machine learning with joint self-adaptive sime mould algorithm. *Comput. Biol. Med.* **2022**, *148*, 105885.
64. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.; et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78.
65. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; et al. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867.
66. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753.
67. Myronenko, A. 3D MRI brain tumor segmentation using autoencoder regularization. In Proceedings of the Brainlesion 2018: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Granada, Spain, 16 September 2018.
68. Liu, Y.; Mu, F.; Shi, Y.; et al. Sf-net: A multi-task model for brain tumor segmentation in multimodal mri via image fusion. *IEEE Signal Process. Lett.* **2022**, *29*, 1799–1803.
69. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; et al. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211.
70. Roth, H.R.; Oda, H.; Zhou, X.; et al. An application of cascaded 3D fully convolutional networks for medical image segmentation. *Comput. Med. Imaging Graph.* **2018**, *66*, 90–99.
71. Wang, F.; Jiang, R.; Zheng, L.; et al. 3d u-net based brain tumor segmentation and survival days prediction. In *International MICCAI Brainlesion Workshop*; Springer: Berlin, Germany, 2019; pp. 131–141.
72. Liu, Z.; Tong, L.; Chen, L.; et al. Canet: Context aware network for brain glioma segmentation. *IEEE Trans. Med. Imaging* **2021**, *40*, 1763–1777.

73. Zhu, Z.; Zhang, Z.; Qi, G.; et al. A dual-branch network for ultrasound image segmentation. *Biomed. Signal Process. Control* **2025**, *103*, 107368.
74. Shi, W.; Xu, J.; Gao, P. Ssformer: A lightweight transformer for semantic segmentation. In Proceedings of the 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), Shanghai, China, 26–28 September 2022.
75. Hatamizadeh, A.; Tang, Y.; Nath, V.; et al. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
76. Zhou, H.Y.; Guo, J.; Zhang, Y.; et al. nnformer: Interleaved transformer for volumetric segmentation. *arXiv* **2021**, arXiv:2109.03201.
77. Hatamizadeh, A.; Nath, V.; Tang, Y.; et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*; Springer: Berlin, Germany, 2021; pp. 272–284.
78. Peiris, H.; Hayat, M.; Chen, Z.; et al. A robust volumetric transformer for accurate 3D tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin, Germany, 2022; pp. 162–172.
79. Lee, S.H.; Lee, S.; Song, B.C. Vision transformer for small-size datasets. *arXiv* **2021**, arXiv:2112.13492.
80. Chandrakar, M.K.; Mishra, A. Brain tumor detection using multipath convolution neural network (CNN). *Int. J. Comput. Vis. Image Process.* **2020**, *10*, 43–53.
81. Zhao, J.; Li, Q.; Li, X.; et al. Automated segmentation of cervical nuclei in pap smear images using deformable multi-path ensemble model. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1514–1518.
82. Wang, N.; Lin, S.; Li, X.; et al. MISSU: 3D medical image segmentation via self-distilling TransUNet. *IEEE Trans. Med. Imaging* **2023**, *42*, 2740–2750.
83. Zhang, Y.; Bai, Z.; You, Y.; et al. Multi-path Feature Fusion and Channel Feature Pyramid for Brain Tumor Segmentation in MRI. In *International Conference on Image and Graphics*; Springer: Berlin, Germany, 2023; pp. 26–36.
84. Zhou, Z.; He, Z.; Jia, Y. AFPNet: A 3D fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images. *Neurocomputing* **2020**, *402*, 235–244.
85. Li, Y.; Cui, W.G.; Huang, H.; et al. Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the Fisher vector approach. *Knowl.-Based Syst.* **2019**, *164*, 96–106.
86. Wang, G.; Li, W.; Ourselin, S.; et al. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In Proceedings of the Brainlesion 2017: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Quebec City, QC, Canada, 14 September 2017.
87. Isensee, F.; Kickingereder, P.; Wick, W.; et al. No new-net. In Proceedings of the Brainlesion 2018: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Granada, Spain, 16 September 2018.
88. Jiang, Z.; Ding, C.; Liu, M.; et al. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In Proceedings of the Brainlesion 2019: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Shenzhen, China, 17 October 2019.
89. Pereira, S.; Pinto, A.; Alves, V.; et al. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1240–1251.
90. Dolz, J.; Gopinath, K.; Yuan, J.; et al. HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE Trans. Med. Imaging* **2018**, *38*, 1116–1126.
91. Liu, Y.; Mu, F.; Shi, Y.; et al. Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion. *Front. Neurosci.* **2022**, *16*, 1000587.
92. Zhang, Y.; Yang, J.; Tian, J.; et al. Modality-aware mutual learning for multi-modal medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021.
93. Mo, S.; Cai, M.; Lin, L.; et al. Multimodal priors guided segmentation of liver lesions in MRI using mutual information based graph co-attention networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020.
94. Zhang, Y.; Li, Z.; Li, H.; et al. Prototype-Driven and Multi-Expert Integrated Multi-Modal MR Brain Tumor Image Segmentation. *IEEE Trans. Instrum. Meas.* **2024**, *74*, 2500614.
95. Li, Z.; Zhang, Y.; Li, H.; et al. Deformation-aware and reconstruction-driven multimodal representation learning for brain tumor segmentation with missing modalities. *Biomed. Signal Process. Control* **2024**, *91*, 106012.
96. Jégou, S.; Drozdzal, M.; Vazquez, D.; et al. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.
97. Gu, Z.; Cheng, J.; Fu, H.; et al. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292.
98. Lee, H.J.; Kim, J.U.; Lee, S.; et al. Structure boundary preserving segmentation for medical image with ambiguous boundary. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4817–4826.

99. Yan, Q.; Liu, S.; Xu, S.; et al. 3D Medical image segmentation using parallel transformers. *Pattern Recognit.* **2023**, *138*, 109432.
100. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
101. He, X.; Qi, G.; Zhu, Z.; et al. Medical image segmentation method based on multi-feature interaction and fusion over cloud computing. *Simul. Model. Pract. Theory* **2023**, *126*, 102769.
102. Baid, U.; Ghodasara, S.; Mohan, S.; et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv* **2021**, arXiv:2107.02314.
103. Xu, Y.; He, X.; Xu, G.; et al. A medical image segmentation method based on multi-dimensional statistical features. *Front. Neurosci.* **2022**, *16*, 1009581.
104. Vijay, S.; Guhan, T.; Srinivasan, K.; et al. MRI brain tumor segmentation using residual Spatial Pyramid Pooling-powered 3D U-Net. *Front. Public Health* **2023**, *11*, 1091850.
105. Kamnitsas, K.; Bai, W.; Ferrante, E.; et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In Proceedings of the Brainlesion 2017: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Quebec City, QC, Canada, 14 September 2017.
106. Kamnitsas, K.; Ferrante, E.; Parisot, S.; et al. DeepMedic for brain tumor segmentation. In Proceedings of the Brainlesion 2016: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Athens, Greece, 17 October 2016.
107. Isensee, F.; Jäger, P.F.; Full, P.M.; et al. nnU-Net for brain tumor segmentation. In Proceedings of the Brainlesion 2020: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Lima, Peru, 4 October 2020.
108. Luu, H.M.; Park, S.H. Extending nn-UNet for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*; Springer: Berlin, Germany, 2021; pp. 173–186.
109. Wu, J.; Fu, R.; Fang, H.; et al. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *Med. Imaging Deep. Learn. PMLR* **2024**, *227*, 1623–1639.
110. Lin, J.; Lin, J.; Lu, C.; et al. CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *IEEE Trans. Med. Imaging* **2023**, *42*, 2451–2461.
111. Zhu, Z.; He, X.; Qi, G.; et al. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf. Fusion* **2023**, *91*, 376–387.
112. Menze, B.H.; Jakab, A.; Bauer, S.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2014**, *34*, 1993–2024.
113. Bakas, S.; Akbari, H.; Sotiras, A.; et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 1–13.
114. Bakas, S.; Reyes, M.; Jakab, A.; et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* **2018**, arXiv:1811.02629.
115. Kinga, D.; Adam, J.B. A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; Volume 5, p. 6.
116. Ho, N.V.; Nguyen, T.; Diep, G.H.; et al. Point-unet: A context-aware point-based neural network for volumetric segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021.
117. Jia, H.; Xia, Y.; Cai, W.; et al. Learning high-resolution and efficient non-local features for brain glioma segmentation in MR images. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020.
118. Zhang, D.; Huang, G.; Zhang, Q.; et al. Exploring task structure for brain tumor segmentation from multi-modality MR images. *IEEE Trans. Image Process.* **2020**, *29*, 9032–9043.
119. Zhou, T.; Ruan, S.; Guo, Y.; et al. A multi-modality fusion network based on attention mechanism for brain tumor segmentation. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, Iowa, USA, 3–7 April 2020; pp. 377–380.
120. Ding, Y.; Yu, X.; Yang, Y. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3975–3984.
121. Zhu, Z.; Yu, K.; Qi, G.; et al. Lightweight medical image segmentation network with multi-scale feature-guided fusion. *Comput. Biol. Med.* **2024**, *182*, 109204.
122. Gessert, N.; Sentker, T.; Madesta, F.; et al. Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. *IEEE Trans. Biomed. Eng.* **2019**, *67*, 495–503.
123. Codella, N.; Rotemberg, V.; Tschandl, P.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368.
124. Singh, V.K.; Kalafi, E.Y.; Wang, S.; et al. Prior wavelet knowledge for multi-modal medical image segmentation using a lightweight neural network with attention guided features. *Expert Syst. Appl.* **2022**, *209*, 118166.

125. Sun, Y.; Dai, D.; Zhang, Q.; et al. MSCA-Net: Multi-scale contextual attention network for skin lesion segmentation. *Pattern Recognit.* **2023**, *139*, 109524.
126. Wu, H.; Chen, S.; Chen, G.; et al. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Med. Image Anal.* **2022**, *76*, 102327.
127. Zhu, Z.; Wang, Z.; Qi, G.; et al. Brain tumor segmentation in MRI with multi-modality spatial information enhancement and boundary shape correction. *Pattern Recognit.* **2024**, *153*, 110553.
128. Qiu, S.; Li, C.; Feng, Y.; et al. GFANet: Gated fusion attention network for skin lesion segmentation. *Comput. Biol. Med.* **2023**, *155*, 106462.
129. Zhao, C.; Lv, W.; Zhang, X.; et al. Mms-net: Multi-level multi-scale feature extraction network for medical image segmentation. *Biomed. Signal Process. Control* **2023**, *86*, 105330.
130. Wang, J.; Huang, G.; Zhong, G.; et al. Qgd-net: A lightweight model utilizing pixels of affinity in feature layer for dermoscopic lesion segmentation. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 5982–5993.
131. Yu, Z.; Yu, L.; Zheng, W.; et al. EIU-Net: Enhanced feature extraction and improved skip connections in U-Net for skin lesion segmentation. *Comput. Biol. Med.* **2023**, *162*, 107081.
132. Iqbal, A.; Sharif, M.; Khan, M.A.; et al. FF-UNet: A U-shaped deep convolutional neural network for multimodal biomedical image segmentation. *Cogn. Comput.* **2022**, *14*, 1287–1302.
133. Valanarasu, J.M.J.; Patel, V.M. Unext: Mlp-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin, Germany, 2022; pp. 23–33.
134. Zhang, X.; Zhang, X.; Ouyang, L.; et al. SMTF: Sparse transformer with multiscale contextual fusion for medical image segmentation. *Biomed. Signal Process. Control* **2024**, *87*, 105458.
135. Song, E.; Zhan, B.; Liu, H. Combining external-latent attention for medical image segmentation. *Neural Netw.* **2024**, *170*, 468–477.
136. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111.
137. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; et al. Kvasir-seg: A segmented polyp dataset. In Proceedings of the MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, Republic of Korea, 5–8 January 2020.
138. Fan, D.P.; Ji, G.P.; Zhou, T.; et al. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; pp. 263–273.
139. Caicedo, J.C.; Goodman, A.; Karhohs, K.W.; et al. Nucleus segmentation across imaging experiments: The 2018 Data Science Bowl. *Nat. Methods* **2019**, *16*, 1247–1253.
140. Han, Z.; Jian, M.; Wang, G.G. ConvUNeXt: An efficient convolution neural network for medical image segmentation. *Knowl.-Based Syst.* **2022**, *253*, 109512.
141. Chen, B.; Liu, Y.; Zhang, Z.; et al. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *8*, 55–68.
142. Zhao, X.; Zhang, L.; Lu, H. Automatic polyp segmentation via multi-scale subtraction network. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021.
143. Dong, B.; Wang, W.; Fan, D.P.; et al. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv* **2021**, arXiv:2108.06932.
144. Bui, N.T.; Hoang, D.H.; Nguyen, Q.T.; et al. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 7985–7994.