Physica Scripta

www.physica.org

IOP Publishing

Physica Scripta

INCLUSIVE PUBLISHING TRUSTED SCIENCE | PURPOSE-LED PUBLISHING™

PAPER

Three-branch dynamic selection network for 3d medical image segmentation

Shumin Ma* and Cao Shi*

Published 15 January 2026 • © 2026 IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Physica Scripta, Volume 101, Number 2

Citation Shumin Ma and Cao Shi 2026 *Phys. Scr.* **101** 026005

DOI 10.1088/1402-4896/ae3441

Authors References Open science

Article information

Abstract

3D medical images present significant challenges for segmentation due to the diversity of anatomical structures, as well as regions with low contrast and noise. To address these issues and achieve precise segmentation, we propose a Three Branch Dynamic Selection Network (TB-Net). This network enhances

Abstract
References

Table of contents for issue 2, volume 101, Physica Scripta

Volume 101
Number 2, January 2026

← Previous issue Next issue →

Open all abstracts

Topical Review

Papers

General and interdisciplinary, including biophysics, geophysics and astrophysics

Educational physics

Retraction

JOURNAL LINKS
Submit an article
About the journal
Editorial Board
Author guidelines
Review for this journal
Publication charges
Awards
Journal collections
Contact us

Three-branch dynamic selection network for 3d medical image segmentation

By Ma, SM (Ma, Shumin) ; Shi, C (Shi, Cao)

View Web of Science ResearcherID and ORCID (provided by Clarivate)

Source PHYSICA SCRIPTA
Volume: 101 Issue: 2
DOI: 10.1088/1402-4896/ae3441

Article Number 026005

Published JAN 16 2026

Indexed 2026-01-20

Document Type Article

Open Peer Reviews ← View open peer reviews

Abstract 3D medical images present significant challenges for segmentation due to the diversity of anatomical structures, as well as regions with low contrast and noise. To address these issues and achieve precise segmentation, we propose a Three Branch Dynamic Selection Network (TB-Net). This network enhances segmentation performance by capturing both rich detailed features and abstract semantic features. TB-Net consists of three parallel feature extraction paths: a CNN branch for capturing local texture and spatial features, a Transformer branch for modeling global semantic information, and a Hidden Connection (HC) branch for learning and reusing historical information from the encoding stage. To efficiently integrate the extracted features, we introduce a Dynamic Selection (DS) module that adaptively fuses multi-branch features. We conducted quantitative experiments on four MRI and CT datasets (OAI-ZIB, Hippocampus, Parse, and Spleen), and TB-Net demonstrated overall superior performance compared with the baseline methods across these datasets. On the OAI-ZIB dataset, TB-Net achieved an average segmentation performance of 90.2% +/- 0.02 Dice, 0.5 +/- 0.1 mm ASSD, and 1.7 +/- 0.4 mm HD95 for femoral cartilage, representing improvements of 0.9% in Dice, a 0.1 mm decrease in ASSD, and a 0.2 mm decrease in HD95 compared with the second-best method.

Keywords **Author Keywords:** 3D medical image segmentation; multi-functional information interaction; hybrid network; feature extraction and reconstruction

Addresses [1] Qingdao Univ Sci & Technol, Coll Informat Sci & Technol, Qingdao, Peoples R China

Categories/ Classification Research Areas: Physics
Citation Topics: 4 Electrical Engineering, Electronics & Computer Science › 4.17 Computer Vision & Graphics › 4.17.128 Deep Visual Recognition
Sustainable Development Goals: 03 Good Health and Well-being

Web of Science Categories Physics, Multidisciplinary

QUERY FORM

JOURNAL: Physica Scripta

AUTHOR: S Ma and C Shi

TITLE: Three-branch dynamic selection network for 3d medical image segmentation

ARTICLE ID: psae3441

Page 1

Q1
Your article has been processed in line with the journal style. Your changes will be reviewed by the Production Editor, and any amendments that do not comply with journal style or grammatical correctness will not be applied and will not appear in the published article.

Page 1

Q2
The layout of this article has not yet been finalized. Therefore this proof may contain columns that are not fully balanced/matched or overlapping text in inline equations; these issues will be resolved once the final corrections have been incorporated.

Page 1

Q3
IMPORTANT: This PDF reference document has been locked to prevent comments. Please use our online Proofing tool to answer all queries, make corrections and add comments.

Page 1

Q4
Please check that **the names, ORCID iDs, and contribution data for all authors are correct**, and that all **authors are linked to the correct affiliations**. Please also confirm that the correct corresponding author has been indicated. **This is your last opportunity to review this information before your article is published.**

Page 1

Q5
The information in the Funding Metadata section is a list of funder names and grant numbers that will be registered on your behalf when the article is published. This information will not be shown in the article text. If an explicit acknowledgment of funding is required, please ensure that it is provided in an Acknowledgments section in your article. Please check that the funding information below is correct for inclusion in the article metadata.
National Natural Science Foundation of China: 62471272, 61806107 , 62201314.

Page 1

Q6
Please specify the corresponding author.

Page 1

Q7
References have been ordered sequentially as per the journal style. Please check and approve.

Page 13

Q8
The author contributions section in your proof uses the information that you provided during submission. To avoid duplication, the contribution information that was in the text of your article has been removed. Please update the data in the proof if required.

Page 13

Q9
A standard data availability statement has been added to your article, based on the information you gave in your submission. This text cannot be changed unless there is an error. Please remove any other data statement if the information is now duplicated.

Page 14

Q10
Please check the details for any journal references that do not have a link as they may contain some incorrect information. If any journal references do not have a link, please update with correct details and supply a Crossref DOI if available.

Page 14

Q11
Please provide the page range or article number in references [6, 30].

Page 14

Q12
Please provide updated details for references [14, 17, 20, 22, 33, 39, 41, 49, 52] if available.

IOP Publishing *Phys. Scr.* **00** (2026) 000000

Physica Scripta

Check for updates

PAPER

Three-branch dynamic selection network for 3d medical image segmentation

Q1 Q2 **RECEIVED**
10 October 2025
Q3 **REVISED**
22 November 2025
Q4 **ACCEPTED FOR PUBLICATION**
6 January 2026
PUBLISHED
DD MM 2026

Shumin Ma and Cao Shi

College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, People's Republic of China

E-mail: shuminma418@qq.com **and** caoshi@yeah.net

Keywords: 3D medical image segmentation, multi-functional information interaction, hybrid network, feature extrac- 20 tion and reconstruction

Abstract

3D medical images present significant challenges for segmentation due to the diversity of anatomical structures, as well as regions with low contrast and noise. To address these issues and achieve precise segmentation, we propose a Three Branch Dynamic Selection Network (TB-Net). This network enhances segmentation performance by capturing both rich detailed features and abstract semantic features. TB-Net consists of three parallel feature extraction paths: a CNN branch for capturing local texture and spatial features, a Transformer branch for modeling global semantic information, and a Hidden Connection (HC) branch for learning and reusing historical information from the encoding stage. To efficiently integrate the extracted features, we introduce a Dynamic Selection (DS) module that adaptively fuses multi-branch features. We conducted quantitative experiments on four MRI and CT datasets (OAI-ZIB, Hippocampus, Parse, and Spleen), and TB-Net demonstrated overall superior performance compared with the baseline methods across these datasets. On the OAI-ZIB dataset, TB-Net achieved an average segmentation performance of 90.2% $\pm$ 0.02 Dice, 0.5 $\pm$ 0.1 mm ASSD, and 1.7 $\pm$ 0.4 mm HD95 for femoral cartilage, representing improvements of 0.9% in Dice, a 0.1 mm decrease in ASSD, and a 0.2 mm decrease in HD95 compared with the second-best method.

Q5

1. Introduction

Q6

Medical image segmentation is crucial in various biomedical applications, playing a significant role in diagnosing diseases, formulating treatment plans, implementing therapies, and quantifying the cancer microenvironment. To alleviate the workload of medical practitioners and reduce the negative impact of subjective factors associated with traditional manual segmentation methods [1, 2], numerous deep learning-based medical image segmentation networks have been developed. These networks are capable of segmenting regions of interest (ROIs) in medical image data from various modalities (such as CT, MRI, etc.) that are of concern to clinicians and biologists [3, 4]. However, the segmentation remains challenging due to issues in three-dimensional medical images, such as edge blurriness and difficulty distinguishing between foreground and background. Therefore, exploring more refined and accurate automatic segmentation methods is necessary.

Over the past few decades, CNNs have made outstanding contributions to the field of computer vision [5–7]. Since the introduction of U-Net [8], the U-shaped architecture has become mainstream in medical image segmentation models. These methods [9–12] consist of convolutional blocks in the encoder to extract features and a decoder to obtain segmentation results. Although CNN-based methods yield good segmentation performance, their inherent locality in convolutional computations shows clear limitations in capturing long-range dependencies. Transformers [13], on the other hand, excel at capturing long-range dependencies. ViT [14] was the first to apply Transformers to image segmentation, achieving remarkable results. Since then, Transformers have been widely adopted in the field of computer vision, leading to the development of methods such as UNETR [15], Swin Transformer [16], and TransUNet [17]. However, due to induction bias,

Transformers require a large amount of labeled data for training, which is challenging in medical imaging tasks. Additionally, Transformers exhibit deficiencies in capturing local features [18], which diminishes the distinguishability between background and foreground.

Recent studies have shown that the combination of CNN and Transformer methods can be more effective [19]. Many hybrid approaches [15, 17, 20–22] have achieved promising results across various tasks. Recent methods such as SegMamba [23] and HiDiff [24] apply the Mamba architecture and diffusion models to image segmentation. However, some existing hybrid methods may not effectively learn the features of either Transformer or CNN. For instance, TransUNet uses CNN as the encoder and only applies Transformer at the lower levels, which may not effectively capture Transformer features. The features from CNN and Transformer are complementary, and their interaction is beneficial for precise segmentation. However, how to accurately embed the features of both still requires further discussion. Additionally, these hybrid methods often overlook the modeling of inter-layer relationships and fail to fully utilize low-level information, which may result in blurred segmentation boundaries.

To alleviate the aforementioned issues, this paper proposes a new three-branch Transformer-CNN hybrid network, featuring dynamic selection and a hidden connection path for various 3D medical image segmentation tasks. To learn Transformer and CNN features, we have designed separate branches for each. To handle dual-branch features and mitigate the weak inductive bias of the Transformer, we introduced a Dynamic Selection Module (DS) that facilitates rich feature interaction. Additionally, to effectively utilize low-level detail information, a hidden connection path (HC) has been designed to retain historical context.

The main contributions of this work are as follows:

- This paper proposes a 3D three-branch interactive network with dynamic selection blocks and hidden connections, which demonstrates good segmentation performance across medical datasets of varying scales.
- To facilitate rich information interaction, we introduce a dynamic fusion module to select information, modeling the information from the dual-branch path to mitigate weak inductive bias.
- We introduce a hidden connection path to transfer historical information to the decoder in order to narrow the semantic gap between the encoder and decoder.
- In four segmentation tasks: OAI-ZIB, Hippocampus, Parse2022, and Spleen, TB-Net outperforms hybrid Transformer-CNN methods or pure CNN approaches, demonstrating good generalization capability.

2. Related work

2.1. Classic CNN or transformer architecture networks

In the initial stages of segmentation research, traditional methods [25, 26] dominated until the emergence of networks such as [8, 11, 27, 28]. The excellent segmentation capabilities of these networks demonstrated the effectiveness of the U-shaped architecture, which has since been widely adopted in the medical field. While pure convolutional networks excel in local feature extraction, they struggle to capture global features effectively. To address this issue, some studies have chosen to directly or indirectly expand the receptive field within CNN architectures. For example, ConvNeXt [12] proposed by Woo *et al* employs a 7×7 convolutional kernel and increases the depth of the network to achieve impressive segmentation results; Yu [29] *et al* proposed atrous convolution, which increases the receptive field by expanding the sampling rate. Hu *et al* introduced GENet [30] and SENet [31], which aggregate global information and compute weights for feature channels through global average pooling to obtain more comprehensive global features. However, simply increasing the size of the convolutional kernel does not always lead to better extraction performance [12]. Additionally, the density of pooling operations increases with the expanded receptive field, which can result in decreased spatial resolution. Inspired by Mednext [32], convolutional blocks with scalable feeler fields were designed to alleviate the above problem, allowing Convs to better play the role of high-pass filters.

With the advent of Transformers, a series of Transformer-based networks has been proposed to address the shortcomings of CNNs. The good segmentation efficiency of ViT [14] demonstrates the effectiveness of Transformers in the visual domain. However, the quadratic computational complexity of Transformer limits its use in three-dimensional scenes. Agent Attention proposed by Han [33] *et al* seamlessly integrates softmax attention and linear attention, achieving a good balance between representational capacity and computational efficiency. However, there are still some redundant computations. The Swin Transformer designed by Liu [16] *et al* utilizes a sliding window mechanism to compute attention only within relevant windows, optimizing performance. Cao [34] *et al* proposed swinunet with unet-like pure transformer based on [16], which is effective on large-scale datasets but has limitations in the field of medical image segmentation where data is scarce. To

achieve better segmentation performance across datasets of varying scales, we employ CNN and Transformer in parallel, leveraging their complementary strengths.

2.2. Hybrid network of CNN and transformer

Hybrid models that combine CNNs and Transformers have been applied in the field of vision and have achieved notable successes. TransUNet, proposed by Chen [17] *et al*, stacks convolutional blocks and Transformers in a serial manner, where the input first passes through convolutional layers and then is processed by Transformers to compensate for the low-resolution features lacking detailed localization information caused by Transformers. Segformer, introduced by Xie [35] *et al*, designs a hierarchical Transformer that does not require positional encoding and utilizes only a few fully connected layers for the decoder. This design enables adaptation to various resolutions. Peng [36] *et al* proposed Conformer, which processes 2D images by designing parallel branches of Transformers and CNNs, using a fusion module to integrate the features from the parallel branches, outperforming individual Transformer or CNN frameworks. Chen [37] *et al* introduced Mobile-Former, an innovative bidirectional cross-bridging method to fuse Transformers and CNNs, resulting in high computational efficiency and strong expressiveness, though it leads to a low parameter sharing rate between the two modules. Hiformer [38], proposed by Heidari *et al* integrates CNN and Transformer on 2D. Although it uses the maximum and minimum level features in feature fusion, there are limitations when reusing features at different scales. In addition, large models such as MedSAM3D [39] and TotalSegmentator [40] have been proposed, which can perform high-precision, automated segmentation of multiple organs and complex structures, significantly enhancing clinical efficiency and reproducibility. Currently, there are few parallel architectures involving both Transformers and CNNs in the field of 3D medical image segmentation, and how to effectively combine the features from both remains a topic worth exploring. According to our experiments (table 6), simple fusion of CNN and Transform branching features yields poor performance, possibly due to the introduction of redundant features. We found that a suitable feature fusion mechanism is crucial for the success of parallel architectures.

2.3. Interaction between the encoder and decoder

In U-Net architectures, the semantic gap between the encoder and decoder may lead to suboptimal segmentation results [41] obtained by the network. Most existing methods compensate for this gap by using skip connections to transmit information from the encoder to the decoder, merging low-level and high-level information to address the loss of features during sampling. However, this simple connection scheme does not effectively model global information [18, 42]. The U-Net 3+ [43] proposed by Huang *et al* attempts to capture multi-granularity features by introducing nested and dense skip connections to fuse high-level and low-level information. Gorkem [44] *et al* introduced DCA, which is not limited by the number of stages and performs double cross-attention calculations on all encoder outputs before connecting them to the corresponding parts of the decoder, thereby narrowing the semantic gap. ACC-UNet [45], proposed by Nabil *et al*, employs a patch-based contextual aggregation method at the skip connections, which is contrary to window attention, to adjust the outputs of multiple stages to the same size, striving to provide the network with rich cross-semantic information. However, these methods lack modeling of sequential relationships and do not reuse historical information, which may lead to segmentations with fuzzy boundaries. Distinguishing from the above methods, we introduce a learnable path bridging encoder-decoder.

3. Method

The architecture of TB-Net is illustrated in figure 1(a), mainly including four parts: dual branching backbone, hidden connection path, dynamic selection module, and decoders. The dual-branch backbone consists of transformer blocks and CNN blocks, which are used to extract transformer features and CNN features, respectively. To integrate the representations extracted from the two branches, we introduce a Dynamic Selection Module (DS). Note that we design a Hidden Connection Path (HC) to model the sequential relationships between levels. In addition, the output information from the dual-branch backbone and the DS block is skip connected to the decoder.

3.1. CNN branch

To capture the details in the image and extract rich local features, we designed the CNN branch. The CNN branch consists of five convolutional layers, each containing Dynamic Receptive Field blocks (DRF) and downsampling blocks with a stride of 2. The structure of the DRF is shown in figure 1(b). The DRF consists of three parallel branches, with each branch including one, two, and three residual blocks to simulate receptive fields of sizes 5, 9, and 13, respectively. The features are then integrated through element-wise addition. [19]

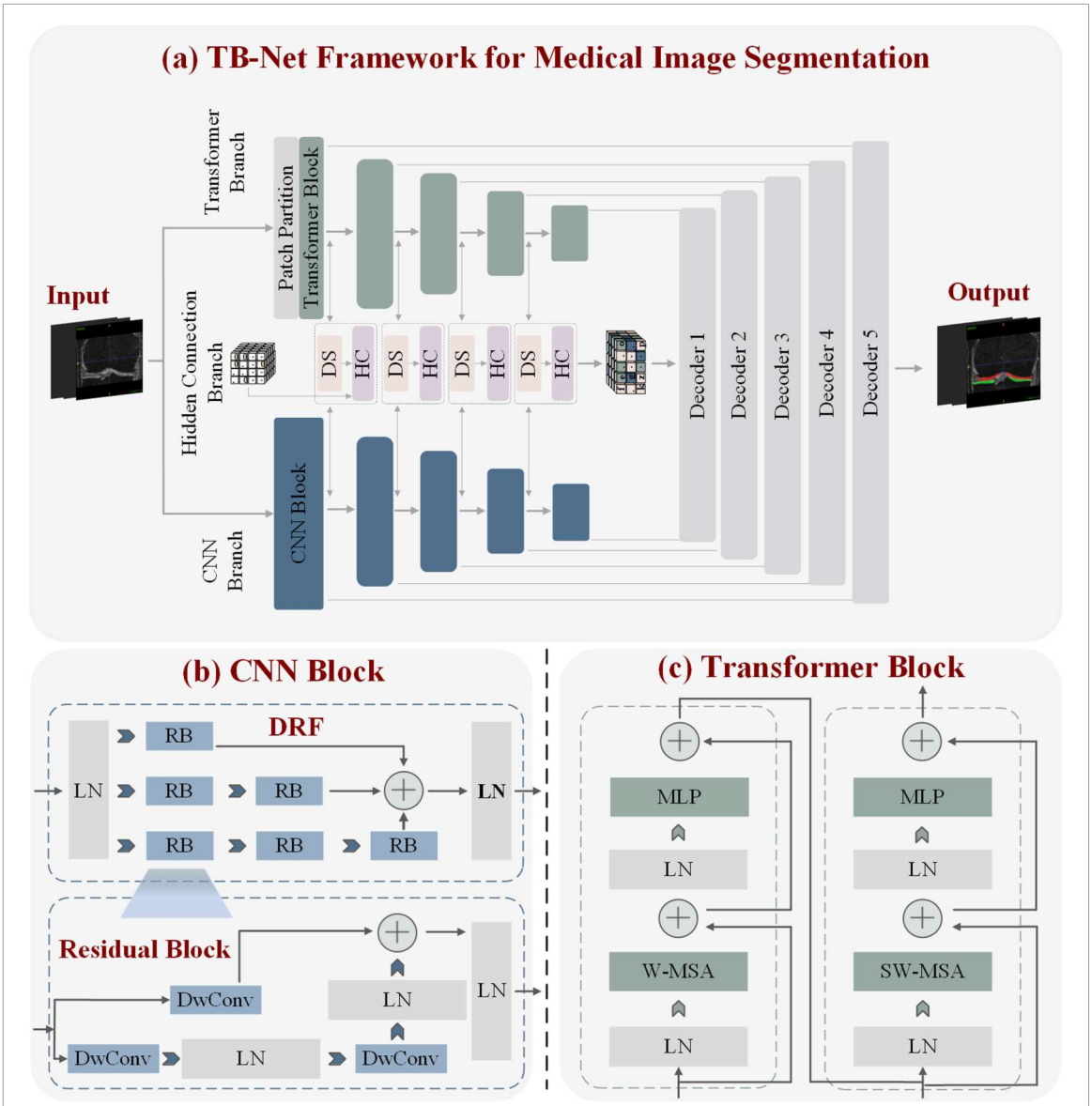

Figure 1. The overall framework of TB-Net. It consists of two parallel feature extraction paths (CNN encoder and Transformer encoder), a parallel hidden connection learning path, and a dynamic selection module. (a) represents the overall structure of the network. (b) shows the components of the CNN encoder. (c) illustrates the Transformer blocks used in the Transformer branch.

pointed out that deep convolutions have an effect similar to local attention, which is why the parallel residual blocks are constructed using depth-wise convolutions. Specifically, to match the dynamic sensory field, we utilize layer normalization in the composition of the DRF instead of batch normalization [46]. This process can be expressed as:

$$RB = \mu(x) \oplus \mu(\mu(x)). \tag{1}$$

$$DRF = \delta(RB_1(X) \oplus RB_2(X) \oplus RB_3(X)). \tag{2}$$

$$F_C^n = DRF(Y). \tag{3}$$

In the formula, x denotes the input feature received by the current branch, Y denotes input, δ indicates the layer normalization(LN), μ refers to the features processed by LN and depth-wise convolution(DWConv), and X denotes the features after LN. $\oplus$ denotes element-wise addition. RB_1 indicates the use of one residual block (RB), with the other two being similar. F_C^n is the feature obtained after DRF, and n represents the layer number where $n = \{1, 2, 3, 4, 5\}$.

3.2. Transformer branch

The Transformer branch is composed of transformers with moving windows, consisting of five stages, each stage comprising 3D Swin-Transformer blocks. These blocks include self-attention mechanisms, sliding window self-attention-based Transformers, as well as LN and MLP layers. As shown in figure 1(c), after the

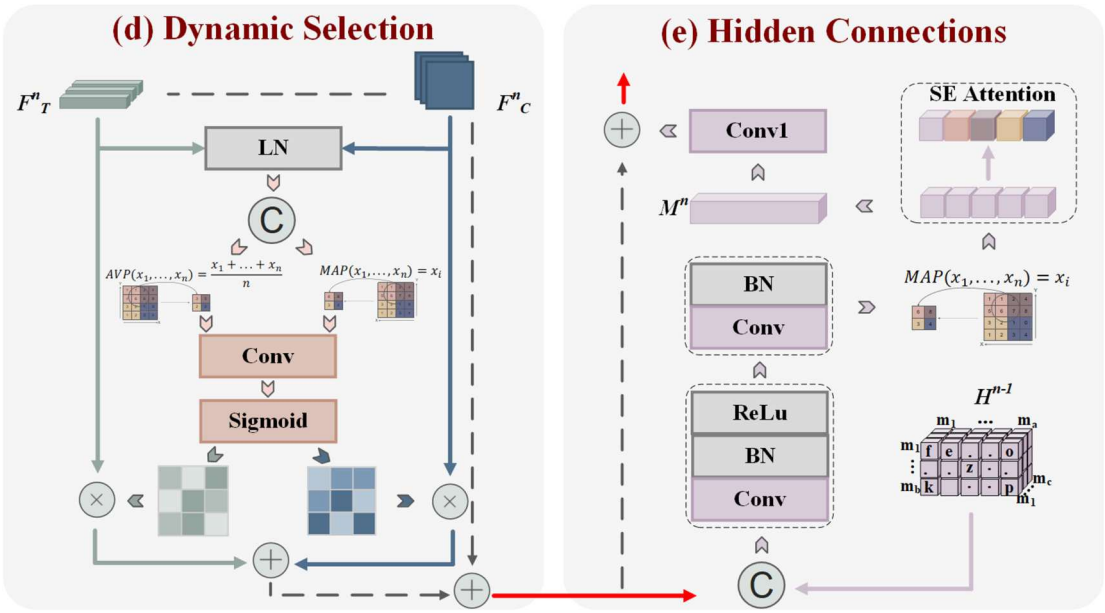

Figure 2. (d) is the structural diagram of the dynamic selection block, which is a fundamental component of the DS. (e) displays the hidden connection block, which is a basic unit of the hidden connection. Here, C denotes the concat operation. F_T^n and F_C^n are the outputs of the n^{th} layer of the CNN and Transformer branches, respectively, while H^{n-1} represent the features of the $(n-1)^{th}$ layer of the hidden connection branch. *Conv*1 is a $1 \times 1 \times 1$ convolution block. The symbol $\otimes$ indicates element-wise multiplication, and $\oplus$ denotes element-wise addition.

patch partition layer of the Transformer branch, input $Y \in \mathbb{R}^{H \times W \times D}$, a 3D image with a spatial resolution of $H \times W \times D$ is divided into patches of size $P_h \times P_w \times P_d$,with dimensions $H/P_h \times W/P_w \times D/P_d \times S$, where S is the dimension of the embedding. In subsequent layers, the input and output are shown in the following equations:

$$F_n^{T'} = W - MSA(\delta(F_{n-1}^T)) + F_{n-1}^T. \tag{4}$$

$$F_n^T = \beta(\delta(F_n^{T'})) + F_n^{T'}. \tag{5}$$

$$F_{n+1}^{T'} = SW - MSA(\delta(F_n^T)) + F_n^T. \tag{6}$$

$$F_{n+1}^T = \beta(\delta(F_{n+1}^{T'})) + F_{n+1}^{T'}. \tag{7}$$

where $F_n^{T'}$ and $F_{n+1}^{T'}$ are the outputs after window multi-head self-attention(W-MSA) and moving window multi-head self-attention(SW-MSA). n represents the layer number where$\{n = 1, 2, 3, 4, 5\}$.δ and β refer to Layer Normalization and Multi-Layer Perceptron.

3.3. Dynamic selection block

To our knowledge, simply concatenating the feature maps from the CNN branch and the transformer branch may lead to misalignment between them. To address this issue, we propose the dynamic selection block(DS) adaptively utilizes contextual information through a large dynamic spatial selection domain. With the dynamic selection block, the features from multiple source branches are embedded, enhancing the availability of complex information. Forcing different branches to directly output features with the same dimension may affect the effective extraction of information. Figure 2(d) illustrates the composition of the DS. Features are aligned using 1x1x1 convolution before being concatenated. Here, F_T^n denotes the features output from the n^{th} layer of the transformer branch, while F_C^n denotes the features output from the n^{th} layer of the CNN branch, which can be formulated as:

$$F = \delta(C(C_1(F_T^n),\ C_1(F_C^n))). \tag{8}$$

Where C_1 represents the $1 \times 1 \times 1$ convolution operation, C denotes the concatenated features, and δ indicates layer normalization, while F is the mixed feature after concatenation. To dynamically select the output features from the two branches, we model the weights λ_1 and λ_2. First, we construct the parameters ω_{MAP} and ω_{AVP} by applying max pooling (MAP) and average pooling (AVP) across the channels. Then, we use a convolutional layer with a kernel size of 7 spatially interact these features, thereby enabling the selection of the extracted features. Finally, a Sigmoid activation function is applied to obtain the dynamic selection weights λ_1 and λ_2. This process is represented as:

$$\lambda_1 = \gamma(C_7(AVP(\omega_{AVP}))). \tag{9}$$

$$\lambda_2 = \gamma(C_7(MAP(\omega_{MAP}))). \tag{10}$$

Here, $\omega_{AVP} = AVP(F)$ and $\omega_{MAP} = MAP(F)$. γ denotes the sigmoid activation function, and C_7 represents a convolutional kernel of size 7. The processing of the DS block at the n^{th} layer can be represented as:

$$F^n = (\lambda_1 \otimes C_1(F_T^n)) \oplus (\lambda_2 \otimes C_1(F_C^n)) + (F_T^n \oplus F_C^n). \tag{11}$$

Where $\oplus$ denotes element-wise addition, $\otimes$ represents element-wise multiplication, and F^n denotes the output of the n^{th} dynamic convolution block($n = \{1, 2, 3, 4\}$). Notably, we also use residual connections within the block.

3.4. Hidden connections

To model the inter-layer relationships, the Hidden Connection Path (HC) is introduced. By enabling the dual backbone to learn shallow detail features, deep abstract semantic features can incorporate historical information. As can be seen in figure 2(e), the input to this path is a zero-initialized learning matrix, which enhances the robustness of the model [15]. The HC consists of n layers, where $n = \{1, 2, 3, 4\}$, and each layer is composed of Hidden Connection Blocks (HCB). The initial size of the zero-initialized matrix is $H \times W \times D$, and each layer downsamples by half while learning historical information through the dual-branch extraction path. The size of the feature map at each layer is $H/2^{n-1} \times W/2^{n-1} \times D/2^{n-1} \times 2^{n-1}C$, where C is the number of input image channels.

Specifically, the output of the Dynamic Selection module (DS) is connected to the output of the previous Hidden Connection Path (HC) layer. The connected features are then passed to two convolutional layers for feature refinement. Max pooling is used to reduce parameters, followed by an attention unit to capture more relevant intermediate layer features M^n. To compress the information, a $1 \times 1 \times 1$ convolutional block processes M^n. The state features H^n of the current layer are obtained through residual connections. This process can be expressed as:

$$M = CL_2(CL_1(C(F^n, H^{n-1}))). \tag{12}$$

$$M^n = \alpha(MAP(M)). \tag{13}$$

$$H^n = C_1(M^n) \oplus H^{n-1}. \tag{14}$$

Where C represents concat, CL_1 consists of a convolution block with a kernel size of 3, followed by batch normalization(BN) and a ReLU activation function. CL_2 contains the same convolution block and BN as CL_1. C_1 is a $1 \times 1 \times 1$ convolution. α represents SE attention, and MAP stands for max pooling. F^n is the output of the DS at the same layer.

3.5. Loss function

We use a loss function which combines the soft Dice loss and cross-entropy loss to accomplish the network's segmentation task. The loss function formulas are as follows:

$$Loss_{Dice} = 1 - \frac{2}{N}\sum_{i=1}^{N} \frac{\sum_{i=1}^{I} g_i p_i + \epsilon}{\sum_{i=1}^{I} g_i + \sum_{i=1}^{I} p_i + \epsilon}. \tag{15}$$

$$Loss_{CE} = -\frac{1}{I} g_i \log p_i + (1 - g_i)\log(1 - p_i). \tag{16}$$

$$Loss_{Total} = Loss_{Dice} + Loss_{CE}. \tag{17}$$

Where N represents the number of classes;I is the number of voxels($H \times W \times D$); g_i denotes the ground truth labels;p_i represents the predicted map, and ϵ is a smoothing factor to prevent division by zero. The network's total loss function is denoted as $Loss_{Total}$.

4. Experiments

4.1. Datasets and evaluation metrics

To quantitatively evaluate the performance of TB-Net, we conducted experiments on four public available datasets: OAI-ZIB, Hippocampus, Parse2022, and Spleen.

4.1.1. OAI-ZIB datasets

OAI-ZIB [47] is a knee joint cartilage dataset provided by the Osteoarthritis Initiative, comprising 507 3D MRI images. The femoral and tibial cartilage in each image was manually annotated by experts. Each image has a size of $384 \times 384 \times 160$, with a voxel spacing of 0.37 mm $\times$ 0.37 mm $\times$ 0.7 mm. The dataset is split into 253 and 254 scans as the training and testing sets.

4.1.2. Hippocampus datasets

The Hippocampus dataset [48] is a header dataset aimed at segmenting two neighboring small structures with high accuracy(anterior and posterior). The dataset is utilized in the Medical Segmentation Decathlon (MSD). It contains MRI images from 394 patients, of which 263are designated for the training set and 131 for the test set. The median image size is $35 \times 50 \times 36$ with a spacing of 1.0 mm $\times$ 1.0 mm $\times$ 1.0 mm.

4.1.3. Spleen datasets

The Spleen dataset [48] is part of the Medical Segmentation Decathlon (MSD). It consists of 61 3D CT scans, with 41 scans for the training set and 20 for the test set. Each image in this dataset has a voxel spacing of 1.6 mm $\times$ 0.79 mm $\times$ 0.79 mm, with an average image size of $187 \times 512 \times 512$.

4.1.4. Parse2022 datasets

The Parse2022 dataset [49](Pulmonary Artery Segmentation), originating from MICCAI 2022, is a CT dataset for pulmonary artery segmentation. It contains 100 publicly available images with corresponding labels. The images have an average size of $512 \times 512 \times 301$, with voxel spacing of 0.6543 mm $\times$ 0.6543 mm $\times$ 1.0 mm.

4.2. Evaluation metrics

To objectively compare the performance of different models, we used average symmetric surface distance (ASSD) [50], dice similarity coefficient (DSC) [51], and 95th percentile Hausdorff Distance (HD95) [52] as evaluation metrics.

4.3. Implementation details

Our network TB-Net is implemented on PyTorch version 1.10.1 and MONAI version 1.3.0. It is trained on one RTX 4090 GPU by the AdamW optimizer with a momentum of 0.9 and initial learning rate of 0.0002. The training was conducted for 300 epochs with batch size of 1. During training, to accommodate memory constraints, the Spleen dataset uses randomly cropped patches of size $32 \times 32 \times 32$ as input to the network, whereas other datasets use patches of size $96 \times 96 \times 96$. The samples are simultaneously normalized to the [0, 1] range and subjected to transformations such as interpolation and random cropping. In addition, the datasets are split into 80% for training, 10% for validation, and 10% for testing. The overlap rate for the sliding window in the Transformer branch is set to 0.5.

5. Results

To validate the performance of our proposed TB-Net, we compared it with several 3D segmentation networks. These networks can be broadly categorized into two types: pure CNN-based architectures and hybrid architectures that combine CNNs with Transformers. This includes purely convolutional networks like V-Net, 3D U-Net, and nnU-Net, as well as CNN-based decoder networks such as Swin UNETR, UNETR, and SegFormer, as well as large models like MedSAM3D. To ensure a fair comparison, all methods except for Swin UNETR were not equipped with pre-trained weights.

5.1. Quantitative comparison on datasets

Table 1 shows the performance of different models in segmenting femoral cartilage (FC) and tibial cartilage (TC). The columns for 'Femoral Cartilage' and 'Tibial Cartilage' display the performance of various models tested on the OAI-ZIB dataset. Quantitative results indicate that our method (TB-Net) surpasses all comparable methods. Specifically, for FC, our proposed method achieves scores of 90.2% and 0.5 mm on the evaluation metrics, Dice Similarity Coefficient (DSC) and Average Symmetric Surface Distance (ASSD), respectively. Compared to the conventional nnU-Net, our method improves the DSC by 1.2%. While the basic nnU-Net has an excellent framework and strategy, it lacks the incorporation of Transformer features. The increase here may be attributed to the rich features extracted by the TB-Net encoder. The second-best model, Swin-UNETR, which is based on Transformer with pre-trained weights, also shows an increase of 0.9% in the DSC metric and a decrease of 0.1 mm in the ASSD metric. In TC segmentation, our method also outperforms both CNN-based and Transformer-based methods, achieving the highest overall DSC and the lowest ASSD. This results in a DSC that is a 1.9% increase compared to 3D U-Net, demonstrating significant improvement. This enhancement is likely due to our method's ability to capture more comprehensive features compared to pure CNN-based approaches. Additionally, the ASSD metric for our method is reduced by 2.7 mm compared to the Transformer-based SegFormer, which may be attributed to the HC's compression of supplementary shallow information, leading to more refined segmentation boundaries. The experimental results demonstrate that our method exhibits superior generalization capability.

Table 1. Results of OAI-ZIB datasets. The best results are highlighted in **bold**. Statistical significance is computed by paired t-tests with Holm–Bonferroni correction, comparing each baseline to ours. '*': $p < 0.05$; '**': $p < 0.01$; '***': $p < 0.001$; n.s.: not significant.

Methods	Femoral cartilage			Tibial cartilage		
	DSC (%)	ASSD (mm)	HD95 (mm)	DSC (%)	ASSD (mm)	HD95 (mm)
V-Net	$87.7 \pm 0.02^{***}$	$1.2 \pm 0.2^{***}$	$1.9 \pm 0.6^{***}$	$84.1 \pm 0.04^{***}$	$1.7 \pm 0.3^{***}$	$2.6 \pm 1.5^{***}$
3D U-Net	$89.0 \pm 0.02^{***}$	$0.8 \pm 0.1^{***}$	$1.9 \pm 0.6^{***}$	$85.0 \pm 0.04^{***}$	$1.0 \pm 0.2^{***}$	$2.4 \pm 1.3^{***}$
nnU-Net	$89.0 \pm 0.02^{***}$	$0.7 \pm 0.1^{***}$	$1.9 \pm 0.5^{***}$	$85.5 \pm 0.04^{***}$	$0.8 \pm 0.3^{***}$	$2.6 \pm 1.5^{***}$
SegFormer	$85.6 \pm 0.02^{***}$	$1.8 \pm 0.1^{***}$	$2.2 \pm 0.4^{***}$	$84.0 \pm 0.04^{***}$	$3.2 \pm 0.2^{***}$	$4.3 \pm 1.3^{***}$
UNETR	$88.3 \pm 0.02^{***}$	$0.7 \pm 0.2^{***}$	$2.0 \pm 0.7^{***}$	$85.3 \pm 0.04^{***}$	$1.0 \pm 0.3^{***}$	$2.6 \pm 1.7^{***}$
Swin-UNETR	$89.3 \pm 0.02^{***}$	$0.6 \pm 0.3^{n.s.}$	$2.4 \pm 1.4^{***}$	$85.3 \pm 0.04^{***}$	$0.8 \pm 0.2^{n.s.}$	$2.9 \pm 1.3^{***}$
Ours	$\mathbf{90.2 \pm 0.02}$	$\mathbf{0.5 \pm 0.1}$	$\mathbf{1.7 \pm 0.4}$	$\mathbf{86.9 \pm 0.03}$	$\mathbf{0.5 \pm 0.3}$	$\mathbf{2.4 \pm 1.5}$

Table 2. Results of hippocampus datasets. The best results are highlighted in **bold**. Statistical significance is computed by paired t-tests with Holm–Bonferroni correction, comparing each baseline to ours. '*': $p < 0.05$; '**': $p < 0.01$; '***': $p < 0.001$; n.s.: not significant.

Methods	Anterior			Posterior		
	DSC (%)	ASSD (mm)	HD95 (mm)	DSC (%)	ASSD (mm)	HD95 (mm)
V-Net	$80.0 \pm 0.2^{***}$	$6.5 \pm 4.4^{***}$	$3.8 \pm 6.8^{***}$	$76.7 \pm 0.2^{***}$	$3.7 \pm 1.9^{***}$	$3.0 \pm 4.8^{***}$
3D U-Net	$88.8 \pm 0.03^{n.s.}$	$1.4 \pm 0.1^{n.s.}$	$1.5 \pm 0.6^{***}$	$87.3 \pm 0.03^{n.s.}$	$1.3 \pm 0.1^{***}$	$1.5 \pm 0.7^{n.s.}$
nnU-Net	$90.0 \pm 0.03^{***}$	$1.5 \pm 0.1^{***}$	$1.4 \pm 0.5^{***}$	$89.0 \pm 0.03^{***}$	$1.5 \pm 0.1^{***}$	$1.5 \pm 0.7^{***}$
SegFormer	$86.0 \pm 0.03^{***}$	$3.0 \pm 0.4^{***}$	$2.1 \pm 3.0^{***}$	$85.0 \pm 0.04^{***}$	$2.1 \pm 0.2^{***}$	$1.7 \pm 0.8^{***}$
UNETR	$88.8 \pm 0.03^{n.s.}$	$1.6 \pm 0.2^{***}$	$1.5 \pm 0.6^{***}$	$87.1 \pm 0.03^{n.s.}$	$1.5 \pm 0.1^{***}$	$1.5 \pm 0.7^{n.s.}$
Swin-UNETR	$89.3 \pm 0.03^{n.s.}$	$1.4 \pm 0.1^{***}$	$1.5 \pm 0.6^{***}$	$87.9 \pm 0.03^{n.s.}$	$1.5 \pm 0.1^{n.s.}$	$1.5 \pm 0.7^{n.s.}$
Ours	$\mathbf{90.2 \pm 0.03}$	$\mathbf{1.2 \pm 0.2}$	$\mathbf{1.3 \pm 0.4}$	$\mathbf{88.3 \pm 0.03}$	$\mathbf{1.3 \pm 0.1}$	$\mathbf{1.4 \pm 0.6}$

Table 3. Results of Spleen datasets. The best results are highlighted in **bold**. Statistical significance is computed by paired t-tests with Holm–Bonferroni correction, comparing each baseline to ours. '*': $p < 0.05$; '**': $p < 0.01$; '***': $p < 0.001$; n.s.: not significant.

Methods	DSC (%)	ASSD (mm)	HD95 (mm)
V-Net	$92.7 \pm 0.05^{n.s.}$	$2.4 \pm 1.5^{n.s.}$	$3.7 \pm 1.7^{n.s.}$
3D U-Net	$94.2 \pm 0.01^{***}$	$0.9 \pm 0.2^{***}$	$1.9 \pm 1.1^{n.s.}$
nnU-Net	$96.0 \pm 0.01^{n.s.}$	$0.6 \pm 0.3^{n.s.}$	$1.0 \pm 0.1^{n.s.}$
SegFormer	$94.2 \pm 0.04^{n.s.}$	$0.9 \pm 0.3^{n.s.}$	$1.4 \pm 0.6^{n.s.}$
UNETR	$94.0 \pm 0.02^{n.s.}$	$0.9 \pm 0.3^{n.s.}$	$1.5 \pm 0.2^{n.s.}$
Swin-UNETR	$94.6 \pm 0.04^{n.s.}$	$0.6 \pm 0.2^{n.s.}$	$1.6 \pm 0.8^{n.s.}$
MedSAM-3D	$67.2 \pm 0.2^{n.s.}$	$2.9 \pm 1.4^{n.s.}$	$10.5 \pm 4.5^{n.s.}$
Ours	$\mathbf{96.2 \pm 0.01}$	$\mathbf{0.9 \pm 0.3}$	$\mathbf{1.1 \pm 0.1}$

In order to validate the generalization capability of the proposed method for the segmentation of different organs, we compared it with six different methods on the Hippocampus dataset. The comparison results are shown in table 2. For the segmentation of the anterior region, our model TB-Net outperformed all methods, achieving the highest overall DSC of 90.2% and the best ASSD of 1.2 mm. Compared to the third-best DSC metric from Swin-UNETR, TB-Net shows outstanding performance with an increase of 1.4%. When compared to the second-best ASSD metric from 3D U-Net, the ASSD is reduced by 0.2 mm. In the segmentation of the posterior region, our proposed method is slightly lower than nnU-Net in DSC, but obtains the optimum in ASSD metric. Compared to UNETR, the DSC improved by 1.2%. In contrast to SegFormer, the ASSD metric decreased by 0.8 mm.

Table 3 presents the results on the Spleen dataset, the Dice score and Average Symmetric Surface Distance of TB-Net are 96.2% and 0.9 mm, respectively. Compared with nnU-Net, our approach has a slight increase of 0.3 mm in the ASSD score, but it achieves the highest DSC score, making it the most accurate segmentation. Relative to the popular Transformer-CNN hybrid method Swin-UNETR, TB-Net shows an improvement of 1.6% in the DSC. In terms of the number of network parameters, the parallel three-branch structure carries a significant amount of information, resulting in a higher parameter count compared to typical convolutional

Table 4. Results of Parse2022 datasets. The best results are highlighted in **bold**. Statistical significance is computed by paired t-tests with Holm–Bonferroni correction, comparing each baseline to ours. '*': $p < 0.05$; '**': $p < 0.01$; '***': $p < 0.001$; n.s.: not significant.

Methods	DSC (%)	ASSD (mm)	HD95 (mm)
V-Net	$82.3 \pm 0.03^{***}$	$0.8 \pm 0.2^{***}$	$5.3 \pm 0.6^{***}$
3D U-Net	$70.0 \pm 0.06^{***}$	$1.3 \pm 0.2^{***}$	$8.1 \pm 1.4^{***}$
nnU-Net	$85.8 \pm 0.04^{n.s.}$	$0.5 \pm 0.1^{***}$	$3.7 \pm 0.6^{n.s.}$
SegFormer	$65.4 \pm 0.05^{***}$	$1.7 \pm 0.3^{***}$	$27.9 \pm 5.4^{***}$
UNETR	$81.0 \pm 0.04^{***}$	$1.0 \pm 0.2^{***}$	$5.8 \pm 0.6^{***}$
Swin-UNETR	$85.1 \pm 0.04^{***}$	$0.5 \pm 0.1^{n.s.}$	$4.2 \pm 0.8^{***}$
MedSAM-3D	$56.0 \pm 0.09^{***}$	$4.6 \pm 1.5^{***}$	$68.7 \pm 16.9^{***}$
Ours	$\mathbf{85.7 \pm 0.04}$	$\mathbf{0.5 \pm 0.1}$	$\mathbf{3.7 \pm 0.6}$

Table 5. Model computational cost.

Methods	Params (M)	FLOPs (G)	Inference time (ms)	GPU Mem (MB)
V-Net	45.60	96.00	29.27	770.62
3D U-Net	21.03	30.20	1.64	14.83
nnU-Net	11.20	250.95	25.42	779.73
SegFormer	16.07	195.14	45.05	2158.75
UNETR	115.18	167.19	30.64	1132.30
Swin-UNETR	63.58	333.64	70.35	1865.00
Ours	132.19	181.52	93.27	1859.64

networks. However, compared to UNETR, our method achieves a 2.2% improvement in DSC with only a slight increase in the number of parameters. Regarding computational complexity, TB-Net is positioned at an average level among methods with Transformer encoders. Table 3 further demonstrates the segmentation comparisons between our network and others, indicating the effectiveness of our network.

Table 4 presents the segmentation results of the models on the PARSE 2022 dataset. Although our method ranks slightly below the best-performing nnU-Net in terms of the Dice coefficient, it achieves comparable performance with nnU-Net on key boundary-accuracy metrics, including ASSD and the 95% Hausdorff distance (3.7 mm), both reaching the optimal level. These results indicate that our method is competitive in terms of overall segmentation accuracy and boundary precision.

Moreover, as shown in table 5, TB-Net (Ours) has 132.19 M parameters and 181.52 G FLOPs, placing it at a medium scale among the four Transformer-based models. Compared to Swin-UNETR (63.58 M parameters, 333.64 G FLOPs) and SegFormer (16.07 M parameters, 195.14 G FLOPs), TB-Net achieves high performance while maintaining an acceptable inference time of 93.27 ms per volume and GPU memory usage of 1859.64 MB, indicating moderate computational cost. In contrast to traditional lightweight CNN models such as 3D U-Net (1.64 ms per volume, but lower Dice than TB-Net), TB-Net offers superior segmentation accuracy and robustness, while its inference efficiency remains sufficient for clinical use. Overall, TB-Net strikes a favorable balance between performance and resource consumption, ensuring high-quality segmentation with practical potential for clinical deployment.

5.2. Qualitative comparison on datasets

Figure 3 qualitatively illustrates the segmentation results of various models on the OAI-ZIB dataset. Each row in figure 3 represents a comparison of segmentation examples from different methods. The purple color indicates femoral cartilage, while green represents tibial cartilage. The first column displays segmentation labels, followed by the segmentation results from TB-Net, 3D U-Net, VNet, nnU-Net, SegFormer, Swin UNETR, and UNETR from left to right. It can be observed that in row (c), our method is more sensitive to irregular shape segmentation, with similar conclusions reflected in row (b). In row (a), the three CNN-based methods excessively smooth the segmentation edges, performing worse than the Transformer-based methods. In our TB-Net model, edge segmentation is improved, showing a slope more similar to the labels.

Figure 4 qualitatively showcases the visual comparison between TB-Net and other comparison models on the Hippocampus dataset. It is worth noting that while all methods achieve coarse-grained segmentation in the Posterior region, our proposed TB-Net excels at capturing fine-grained edge features, resulting in impressive segmentation performance. These qualitative and quantitative results highlight the excellent performance of the proposed TB-Net in the Hippocampus segmentation task.

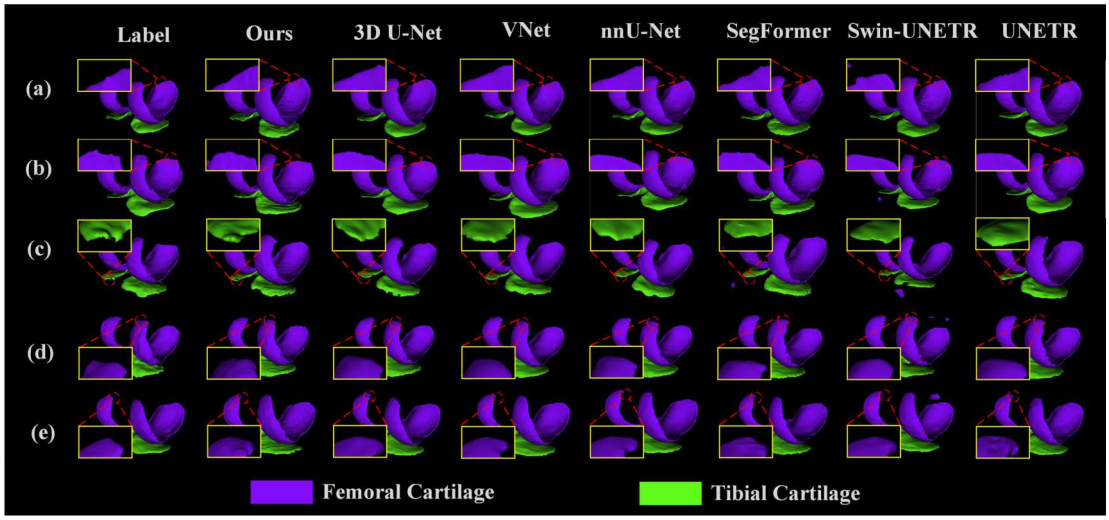

Figure 3. Visualization of the representative SOTAs segmentation results from five methods on the OAI-ZIB dataset. The purple and green areas represent the femoral cartilage (FC) and tibial cartilage (TC), respectively.

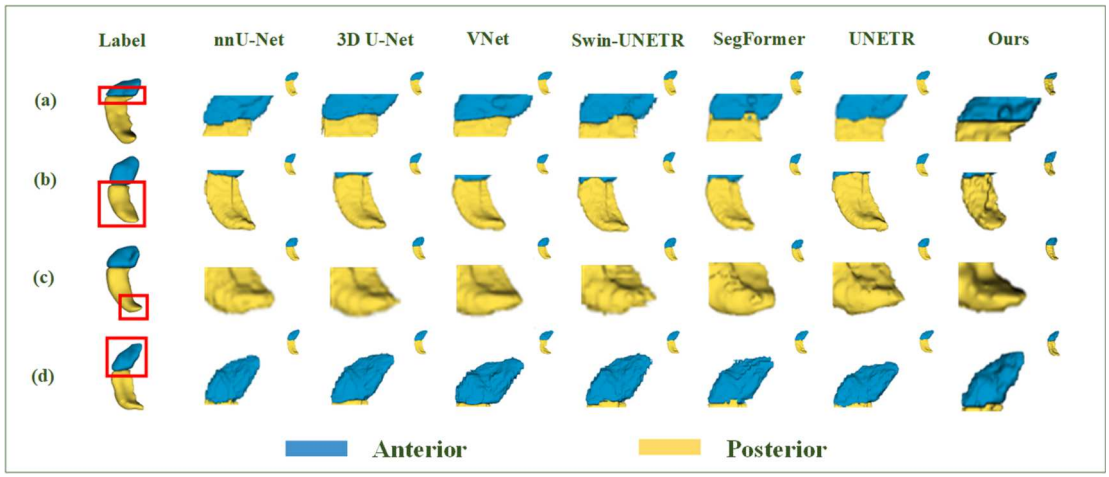

Figure 4. Visualization of the representative SOTAs segmentation results from four methods on the Hippocampus dataset. The blue and yellow areas represent the Anterior and Posterior regions, respectively.

Figure 5 displays the visual segmentation results of seven methods on the Spleen dataset. The Spleen dataset has a limited number of samples, making it challenging to train Transformer-based networks. However, as seen in row (c), our method still outperforms three CNN-based networks in segmenting irregular edges. Additionally, TB-Net's segmentation results for the irregular parts of the organ are closer to the ground truth labels, as shown in rows (d) and (e).

We performed Grad-CAM heatmap visualizations on the same samples from the OAI-ZIB and Spleen datasets, generating results at the corresponding encoder outputs for each branch. From the sagittal slices in figure 6(a), each branch focuses on distinct regions, with the CNN branch providing complementary information to the Transformer branch. Similarly, the HC branch shows attention to edge regions, as illustrated in the third row of the sagittal slice in figure 6(b).

5.3. Ablation study

To evaluate the effectiveness of the several modules we proposed in the network, we conducted multiple experiments on the OAI-ZIB, Hippocampus, and Spleen dataset. The baseline model consists of an encoder constructed from a Swin Transformer and a decoder made up of convolutional blocks. We then added a CNN branch to the baseline model, referred to as Model A. Next, we incorporated the DS module into Model A, calling it Model B. By adding the HC module to Model A, we create Model C. Finally, by adding the HC module to Model B (i.e., TB-Net), we obtain Model D. table 6 presents the experimental results of the various models.

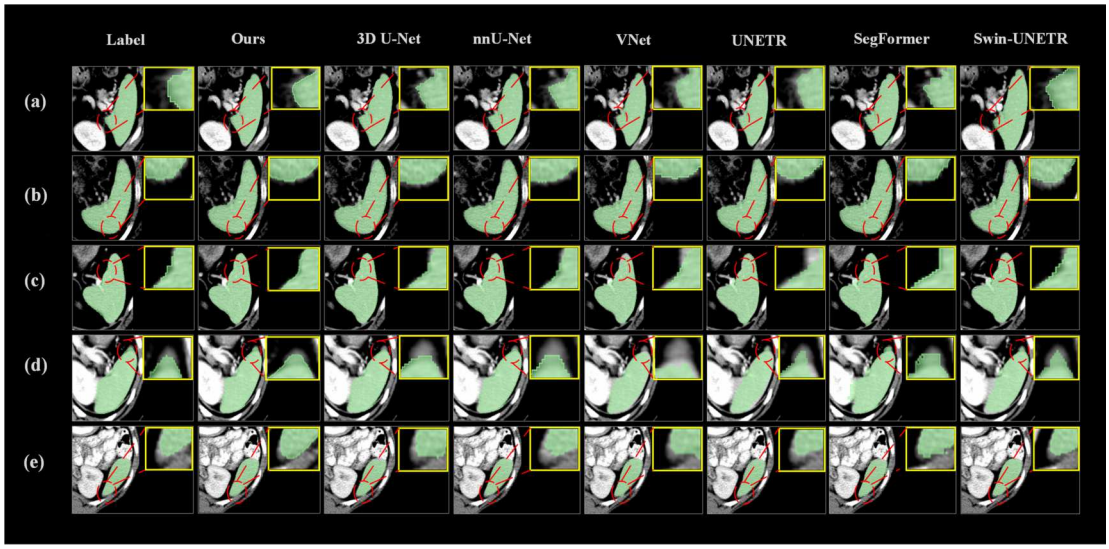

Figure 5. Visualization of the representative SOTAs segmentation results for five patients in the Spleen dataset. The green area represents the segmentation results, and the yellow box contains the enlarged detail view.

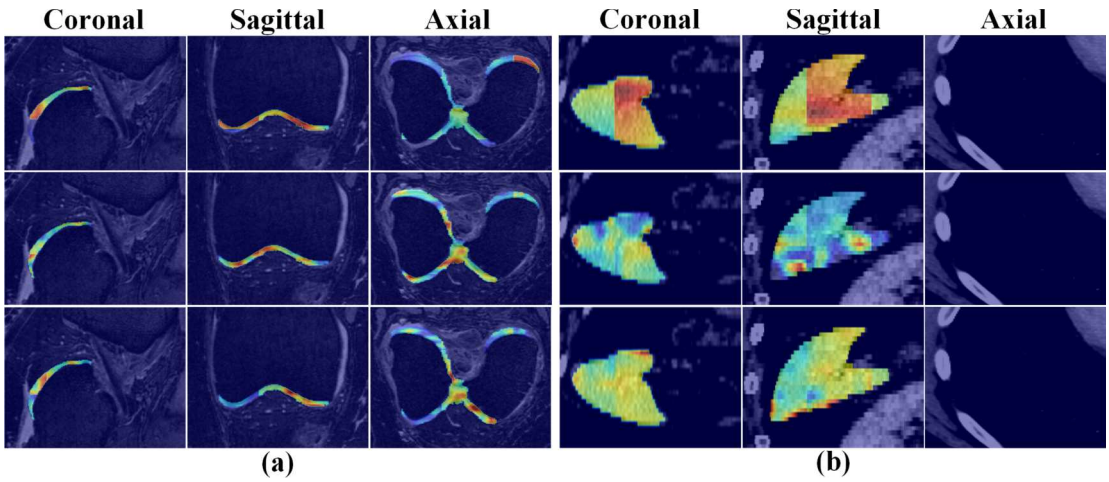

Figure 6. Visualization Heatmaps of Each Branch. The first row depicts visualizations for the Transformer branch, the second row corresponds to the CNN branch, and the third row represents the HC branch. (a) shows results for the OAI-ZIB dataset, and (b) shows results for the Spleen dataset. In the heatmaps, red and yellow indicate regions of high model attention.

Table 6. Ablation study on the spleen dataset.

Methods	DSC(%)	ASSD(mm)
baseline	94.2 ± 0.01	0.9 ± 0.3
Model A	93.7 ± 0.01	1.1 ± 0.3
Model B	94.3 ± 0.01	1.4 ± 0.3
Model C	94.5 ± 0.01	1.2 ± 0.3
Model D	**96.2 ± 0.01**	**0.9 ± 0.3**

5.3.1. Experimental results with dual branches

To validate the effectiveness of the parallel encoder, we conducted experiments. The results on the OAI-ZIB dataset are shown in table 7. Model A, which only adds a CNN auxiliary branch, did not improve the DSC metric, and the ASSD remained unchanged, indicating that simply adding local convolutional information is insufficient to significantly enhance segmentation performance. A similar trend was observed on the Hippocampus dataset, as shown in table 8. Model A did not improve performance. Subsequent experimental results demonstrate the necessity of introducing the DS module for information filtering.

Table 7. Ablation study on the OAI-ZIB dataset.

Methods	Femoral cartilage		Tibial cartilage	
	DSC (%)	ASSD (mm)	DSC (%)	ASSD (mm)
baseline	89.9 ± 0.02	0.5 ± 0.01	85.7 ± 0.02	0.5 ± 0.03
Model A	89.6 ± 0.02	0.5 ± 0.01	85.1 ± 0.03	0.6 ± 0.03
Model B	90.0 ± 0.02	0.7 ± 0.01	85.8 ± 0.03	0.5 ± 0.03
Model C	90.1 ± 0.02	0.6 ± 0.01	86.0 ± 0.03	0.6 ± 0.03
Model D	**90.2 ± 0.02**	**0.5 ± 0.01**	**86.9 ± 0.03**	**0.5 ± 0.03**

Table 8. Ablation study on the Hippocampus dataset.

Methods	Anterior		Posterior	
	DSC (%)	ASSD (mm)	DSC (%)	ASSD (mm)
baseline	89.5 ± 0.03	1.3 ± 0.02	87.3 ± 0.03	1.4 ± 0.01
Model A	88.9 ± 0.03	1.5 ± 0.02	86.8 ± 0.03	1.6 ± 0.01
Model B	89.8 ± 0.03	1.7 ± 0.02	87.8 ± 0.03	1.6 ± 0.01
Model C	89.7 ± 0.03	1.6 ± 0.02	87.6 ± 0.03	1.9 ± 0.01
Model D	**90.2 ± 0.03**	**1.2 ± 0.02**	**88.3 ± 0.03**	**1.3 ± 0.01**

5.3.2. Effectiveness of the DS module

Table 6 indicates that the DS module helps to improve performance. When the DS module is equipped alone, the DSC increases by 0.1% (from 94.2% → 94.3%), while the model's parameter count and computational complexity remain largely unchanged. The DS module effectively emphasizes key features when faced with different input data, enhancing segmentation performance. This may explain why Model B has a higher DSC than Model A. This further underscores the importance of extracting rich features for segmentation accuracy.

5.3.3. Effectiveness of the HC module

To demonstrate the effectiveness of the HC module, we equipped it independently in Model A. Table 6 presents the experimental results on the Spleen dataset. It can be noted that without the HC module, the baseline DSC metric decreased by 0.3% compared to when it is included. In terms of the ASSD metric, the baseline improved by 0.3 mm. This may be because the information compressed by the HC module, when fused with the dual-branch backbone, generates redundant information in the absence of the DS module, thereby affecting the segmentation boundaries, which confirms the effectiveness of the DS module. Furthermore, the enhancement in segmentation performance of Model C may be attributed to the historical information retained by the HC module.

5.3.4. Effectiveness of the dual branches, DS module, and HC module

As shown in table 6, when the DS module and HC module are simultaneously equipped on Model A, both the DSC and ASSD metrics show significant improvement. Compared to the methods with Model B, the maximum improvement reaches 2.1%. This indicates that the dual-branch backbone, DS module, and HC module can mutually enhance each other. Experimental results demonstrate that the DS module can filter key features, while the HC module can compress effective historical information to supplement the detail information required for segmentation.

Furthermore, on the OAI-ZIB and Spleen datasets, we visualized Grad-CAM heatmaps for the same samples with and without the HC module, generated at the outputs of the same-layer encoders in both branches. From the sagittal slices in the first and second rows of figure 7(a), it can be seen that the Trans branch exhibits a more complete response to the target region after incorporating the HC module. Similarly, the CNN branch shows a broader distribution of attention regions, as illustrated in the axial slices in the third and fourth rows of figure 7(a).

5.4. Discussion

In this paper, we propose a three-branch dynamic selection network that integrates historical information for 3D medical image segmentation. Our method achieves accurate segmentation on multiple datasets and outperforms comparative approaches. TB-Net introduces a dynamic feature selection mechanism that adaptively balances contextual features, structural details, and historical information across three

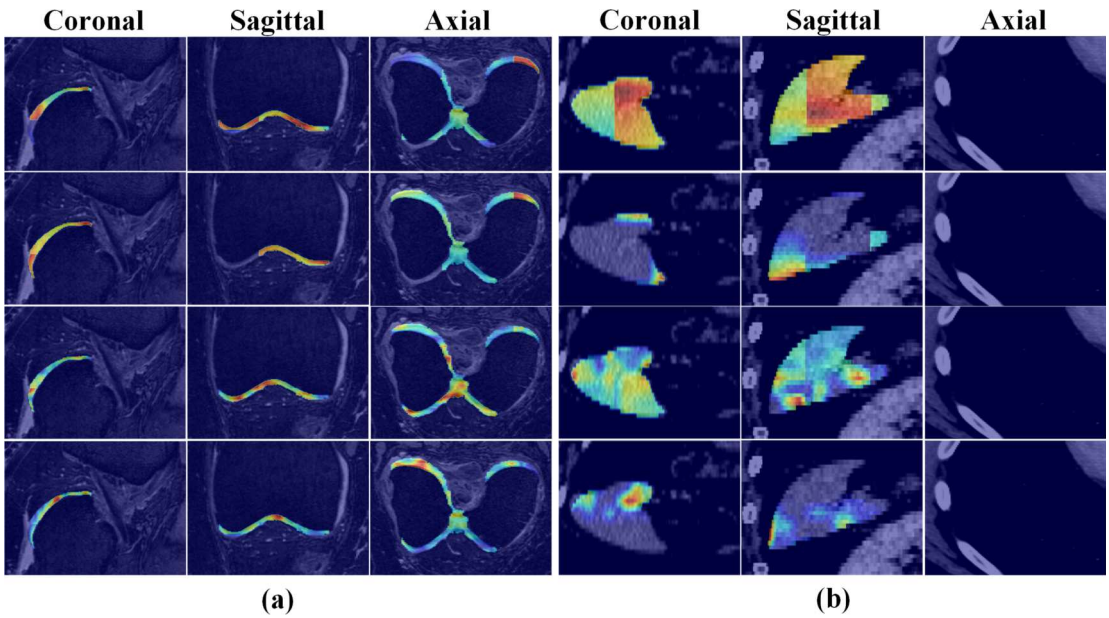

Figure 7. Visualization of HC Module Ablation. The first and second rows display the visualizations of the Transformer (Trans) branch with and without the HC module, respectively, while the third and fourth rows correspond to the CNN branch. (a) shows results for the OAI-ZIB dataset, and (b) for the Spleen dataset. In the heatmaps, red and yellow regions indicate areas of high model attention.

complementary branches. This design enables the model to better handle complex structures and regions with blurred boundaries. However, this study still has some limitations. First, the multi-branch structure increases computational cost and memory consumption, which may limit large-scale clinical deployment. Second, the experiments were conducted on a limited number of datasets, and further validation on more diverse types of medical images is needed. Future work will focus on developing lightweight solutions and extending the multi-branch framework to other medical segmentation tasks to enhance model interpretability and clinical reliability.

6. Conclusion

In this paper, we have proposed a three-branch dynamic selection network incorporating historical information for 3D medical image segmentation. Our method achieved accurate segmentation on four datasets, outperforming several comparison methods. Future research needs to optimize the computational complexity of this method and further develop lightweight approaches.

Q8

Funding

Sample text inserted for demonstration. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62471272, 61806107 and 62201314, the Opening Project of State Key Laboratory of Digital Publishing Technology.

Q9

Data availability statement

The data cannot be made publicly available upon publication because they are not available in a format that is sufficiently accessible or reusable by other researchers. The data that support the findings of this study are available upon reasonable request from the authors.

Author contributions

Shumin Ma
Conceptualization (equal), Methodology (equal), Software (equal), Visualization (equal), Writing – original

draft (equal), Writing – review & editing (equal)

References

[1] Chen X, Williams B M, Vallabhaneni S R, Czanner G, Williams R and Zheng Y 2019 Learning active contour models for medical image segmentation *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 11632–40

[2] Chen B, Liu Y, Zhang Z, Lu G and Kong A W K 2024 TransAttUnet: multi-level attention-guided U-Net with transformer for medical image segmentation *IEEE Transactions on Emerging Topics in Computational Intelligence* **8** 55–68

[3] Xu C, Li Y, Shi C, Zhang H, Bi H and Chen Y 2023 HiM: hierarchical multimodal network for document layout analysis *Applied Intelligence* **53** 24314–26

[4] Shi C, Xu C, He J, Chen Y, Cheng Y, Yang Q and Qiu H 2022 Graph-based convolution feature aggregation for retinal vessel segmentation *Simul. Modell. Pract. Theory* **121** 102653

[5] Kaiming H *et al* 2016 Deep residual learning for image recognition *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 34 pp 770–8

[6] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* **25**

[7] Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T and Xie S 2022 A convnet for the 2020s *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 976–11

[8] Ronneberger O, Fischer P and Brox T U-net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (Springer) 2015 pp 234–41

[9] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N and Liang J Unet++: a nested u-net architecture for medical image segmentation *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* (Springer) 2018 pp 3–11

[10] Çiçek Ö., Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O 3d u-net: learning dense volumetric segmentation from sparse annotation *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19* (Springer) 2016 pp 424–32

[11] Milletari F, Navab N and Ahmadi S-A 2016 V-net: fully convolutional neural networks for volumetric medical image segmentation *Fourth International Conference on 3D Vision (3DV)* (IEEE) pp 565–71

[12] Woo S, Debnath S, Hu R, Chen X, Liu Z, Kweon I S and Xie S 2023 Convnext v2: co-designing and scaling convnets with masked autoencoders *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 133–16

[13] Ashish V 2017 Attention is all you need *Advances in Neural Information Processing Systems* **30** I

[14] Dosovitskiy A *et al* 2020 An image is worth 16x16 words: transformers for image recognition at scale arXiv:2010.11929

[15] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth H R and Xu D 2022 Unetr: transformers for 3d medical image segmentation *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* pp 574–84

[16] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B 2021 Swin transformer: hierarchical vision transformer using shifted windows *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp 10012–22

[17] Chen J, Lu Y and Yu Q T 2021 Transformers make strong encoders for medical image segmentation arXiv:2102.04306

[18] Peng Z, Huang W, Gu S, Xie L, Wang Y, Jiao J and Ye Q 2021 Conformer: local features coupling global representations for visual recognition *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp 367–76

[19] Jiang M, Khorram S and Fuxin L 2024 Comparing the decision-making mechanisms by transformers and cnns via explanation methods *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 9546–55

[20] Chang Y, Menghan H, Guangtao Z and Xiao-Ping Z 2021 Transclaw U-net: Claw U-net with transformers for medical image segmentation arXiv:2107.05188

[21] Xie Y, Zhang J, Shen C and Xia Y 2021 Cotr: efficiently bridging cnn and transformer for 3d medical image segmentation *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24* (Springer) pp 171–80

[22] Zhou H-Y, Guo J, Zhang Y, Yu L, Wang L and Yu Y 2021 nnformer: interleaved transformer for volumetric segmentation arXiv:2109.03201

[23] Xing Z, Ye T, Yang Y, Liu G and Zhu L 2024 SegMamba: long-range sequential modeling mamba for 3D medical image segmentation *Medical Image Computing and Computer Assisted Intervention—MICCAI* ed M G Linguraru *et al (Cham)* (Springer Nature Switzerland) pp 578–88

[24] Chen T, Wang C, Chen Z, Lei Y and Shan H 2024 HiDiff: hybrid diffusion framework for medical image segmentation *IEEE Trans. Med. Imaging* **43** 3570–83

[25] Xu C, Shi C, Bi H, Liu C, Yuan Y, Guo H and Chen Y 2021 A page object detection method based on mask r-cnn *IEEE Access* **9** 448–143

[26] Tsai A, Yezzi A, Wells W, Tempany C, Tucker D, Fan A, Grimson W E and Willsky A 2003 A shape-based approach to the segmentation of medical imagery using level sets *IEEE Trans. Med. Imaging* **22** 137–54

[27] Xiao X, Lian S, Luo Z and Li S 2018 Weighted res-unet for high-quality retina vessel segmentation *9th International Conference on Information Technology in Medicine and Education (ITME)* (IEEE) pp 327–31

[28] Cai S, Tian Y, Lui H, Zeng H, Wu Y and Chen G 2020 Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network *Quantitative Imaging in Medicine and Surgery* **10** 1275

[29] Yu F, Koltun V and Funkhouser T 2017 Dilated residual networks *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 472–80

[30] Hu J, Shen L, Albanie S, Sun G and Vedaldi A 2018 Gather-excite: exploiting feature context in convolutional neural networks *Advances in Neural Information Processing Systems* **31**

[31] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 7132–41

[32] Roy S, Koehler G, Ulrich C, Baumgartner M, Petersen J, Isensee F, Jaeger P F and Maier-Hein K H 2023 Mednext: transformer-driven scaling of convnets for medical image segmentation *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer) pp 405–15

[33] Han D, Ye T, Han Y, Xia Z, Song S and Huang G 2023 Agent attention: on the integration of softmax and linear attention arXiv:2312.08874
[34] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q and Wang M 2022 Swin-unet: Unet-like pure transformer for medical image segmentation *European Conference on Computer Vision* (Springer) pp 205–18
[35] Xie E, Wang W, Yu Z, Anandkumar A, Alvarez J M and Luo P 2021 Segformer: simple and efficient design for semantic segmentation with transformers *Advances in Neural Information Processing Systems* **34** 12 077 12 090
[36] Chen Y, Dai X, Chen D, Liu M, Dong X, Yuan L and Liu Z 2022 Mobile-former: bridging mobilenet and transformer *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 5270–9
[37] Shi C, Xu C, Bi H, Cheng Y, Li Y and Zhang H 2022 Lateral feature enhancement network for page object detection *IEEE Trans. Instrum. Meas.* **71** 1–10
[38] Heidari M, Kazerouni A, Soltany M, Azad R, Aghdam E K, Cohen-Adad J and Merhof D 2023 Hiformer: hierarchical multi-scale representations using transformers for medical image segmentation *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* pp 6202–12
[39] Wang H *et al* 2024 SAM-Med3D: towards general-purpose segmentation models for volumetric medical images arXiv:2310.15161
[40] Wasserthal J *et al* 2023 TotalSegmentator: robust segmentation of 104 anatomic structures in CT images *Radiology: Artificial Intelligence* **5** e230024
[41] Han Q, Fan Z, Dai Q, Sun L, Cheng M-M, Liu J and Wang J 2021 On the connection between local attention and dynamic depth-wise convolution arXiv:2106.04263
[42] Xu C-h, Shi C and Chen Y-n 2021 End-to-end dilated convolution network for document image semantic segmentation *Journal of Central South University* **28** 1765–74
[43] Ates G C, Mohan P and Celik E 2023 Dual cross-attention for medical image segmentation *Eng. Appl. Artif. Intell.* **126** 107139
[44] Ibtehaz N and Kihara D 2023 Acc-unet: a completely convolutional unet model for the 2020s *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer) pp 692–702
[45] Pang Y, Li Y, Shen J and Shao L 2019 Towards bridging semantic gap to improve semantic segmentation *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp 4230–9
[46] Wei H, Feng L, Chen X and B. A 2020 Combating noisy labels by agreement: a joint training method with co-regularization *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 13726 13735
[47] Ambellan F, Tack A, Ehlke M and Zachow S 2019 Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: data from the osteoarthritis initiative *Med. Image Anal.* **52** 109–18
[48] Antonelli M *et al* 2022 The medical segmentation decathlon *Nat. Commun.* **13** 4128
[49] Luo G *et al* 2023 Efficient automatic segmentation for multi-level pulmonary arteries: the PARSE challenge arXiv:2304.03708
[50] Yeghiazaryan V and Voiculescu I 2018 Family of boundary overlap metrics for the evaluation of medical image segmentation *Journal of Medical Imaging (Bellingham)* **5** 015006
[51] Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R and Blaschko M B 2020 Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index *IEEE Trans. Med. Imaging* **39** 3679–90
[52] Celaya A, Riviere B and Fuentes D 2024 A generalized surface loss for reducing the hausdorff distance in medical imaging segmentation arXiv:2302.03868