Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on image segmentation techniques based on algebraic topology methods in computer vision

Guanjie Yan[1], Mengchen Ma[2,✉], Fahui Miao[3]

[1] *The Department of Basic Education, Shanghai Urban Construction Vocational College, Shanghai, 201499, China*
[2] *The Department of Basic Education, Shanghai Jiguang Polytechnic College, Shanghai, 201901, China*
[3] *College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China*

ABSTRACT
Medical image segmentation is the basis for realizing intelligent medical treatment, and plays a very important clinical significance in the localization and identification of lesion areas and the formulation of surgical plans. In this paper, we investigate the image segmentation techniques based on algebraic topology methods in computer vision, and propose an image segmentation network model based on asymmetric topology preservation (ATSNet), with a view to applying it to clinical practice.The ATSNet model adopts the parallel branching structure of CNN and Transformer in the coding part, and proposes a hybrid feature aggregation strategy (HFAS) to achieve image segmentation with high efficiency. Comparison experiments on three benchmark datasets and one clinical dataset prove that the ATSNet model proposed in this paper achieves better results on different datasets, and the statistical analysis results obtained by the model are consistent with those of clinical experts (P>0.05). Meanwhile, ablation experiments demonstrate the effectiveness of the hybrid feature aggregation strategy used in this paper in improving the image segmentation performance of the model. In addition, the proposed method in the Transformer branch when the number of network layers is 3 when the overall accuracy of the largest, and the use of bilateral filtering can be better edge retention, improve the effect of image segmentation. This paper provides a technical path for the practical application of image segmentation technology.

Keywords: algebraic topology, image segmentation, CNN, Transformer

1. Introduction

Computer vision refers to the simulation of biological vision with computers and related equipment. That is, by processing the collected pictures and videos to obtain the corresponding information, it can realize the functions of object recognition, shape and orientation confirmation, motion judgment, etc.,

✉ Corresponding author.
E-mail address: mengchen_ma666@163.con (M. Ma).
Received 03 October 2024; Revised 20 December 2024; Accepted 10 February 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-024

in order to adapt to and understand the external environment and control its own motion [1-3]. In short, computer vision aims to study how to make machines learn to "see", which is the extension of biological vision on machines [4]. Computer vision integrates a number of disciplines such as computer science and engineering, signal processing, physics, applied mathematics and statistics, and involves a number of technologies such as image processing, pattern recognition, artificial intelligence, signal processing and so on [5-6]. Especially with the help of deep learning, computer vision technology performance has made important breakthroughs, becoming one of the basic application technologies of artificial intelligence, which is a necessary means to realize automation and intelligence. Computer vision technology is inherited from image processing and machine vision technology, but the three are different [7-9]. Image processing is mainly based on the color, shape, size and other basic features of digital images to process images [10]. Machine vision, on the other hand, measures and judges target morphological information through machine vision products instead of human eyes. Computer vision, on the other hand, usually includes the image processing process with additional functions such as pattern recognition, which focuses on perception and recognition compared to machine vision that focuses on precise geometric measurement calculations [11-12].

With the wide coverage of artificial intelligence applications in various fields, computer vision technology, which is one of the important branches in this field, has also been rapidly developed [13]. Computer vision technology mainly uses media instruments to obtain relevant information from still images or recorded videos, and then integrates this information through algorithms to realize the understanding of the scenes in pictures or videos [14-15]. In particular, image semantic segmentation has become a key technique in many applications because it provides very useful structured information by taking into account both high-level semantic information and low-level pixel position information, including, for example, the analysis of the surrounding environment in automated driving systems, the beautification of localized areas in cell phone cameras, the localization of irregular tissues or tumors in medical imaging devices, the analysis of the flow of people and the aggregation of people in crowd monitoring systems, and satellite image analysis for satellite imaging systems, etc. [16-17]. Therefore, designing a robust and highly generalized algorithm to solve the image semantic segmentation problem in different scenarios has become an important research direction in the field of computer vision.

Image semantic segmentation is an image pixel classification technique developed in the 20th century, which, unlike the traditional image segmentation problem, requires determining the target label of each pixel in an image, and thus is a pixel-level image understanding task. Since image processing and image analysis are the most important research topics in the field of computer vision, and image semantic segmentation is a fundamental task for scene understanding, it is usually defined with automatic image annotation as the two key tasks for large-scale image processing and understanding problems [18-20]. As more and more application scenarios demand accurate and efficient segmentation, especially in scene understanding, we must know the specific content in the scene in order to integrate the obtained information, which needs to be obtained by image semantic segmentation methods [21].

As an image classification technique, the implementation methods of image semantic segmentation can be divided into three stages chronologically [22]. In the first stage, researchers used simple image segmentation methods for processing some simpler grayscale images, and these algorithms assumed that image pixels of the same category had similar grayscale values, and obtained good results in early segmentation tasks [23-24]. With the rapid development of computer vision technology in recent years, people need to extract more complex color images, as well as multi-channel remote sensing images, and such images no longer meet the data assumptions of the first stage, resulting in simple image segmentation algorithms can not be a good solution to the existing problems [25-26]. For this reason, research scholars have proposed a segmentation method based on the graph model, which transforms the image semantic segmentation problem into a graph cut problem, and greatly improves

the segmentation accuracy by obtaining a certain amount of data information from the dataset through the distance function [27]. Since then, since deep learning has become a big hit in the field of computer vision, image semantic segmentation has also ushered in the third stage of its development. In this stage, the performance of image semantic segmentation has made a qualitative leap and started to move from academia to the commercial world, and gradually shifted from coarse-grained semantic segmentation to fine-grained semantic segmentation [28-29]. At present, due to the great success of deep learning, the early semantic segmentation methods are also fading out of the stage, and how to apply deep learning more fully to image semantic segmentation has become the biggest challenge and the biggest opportunity in the field of image semantic segmentation at this stage.

Scholars mainly use the literature review method to analyze and summarize the image segmentation technology, which involves topics such as the advantages and limitations of image segmentation technology empowered by various deep learning algorithms, the performance of deep learning architectures, and the development trend of future research on image segmentation technology. Literature [30] categorizes and analyzes image segmentation techniques based on different deep learning algorithms to understand the advantages and disadvantages of image segmentation techniques empowered by different deep learning algorithms as well as the applicable places. Literature [31] systematically reviews the research on semantic and instance segmentation for computer vision image segmentation, compares the performance of various image segmentation algorithms, and also looks into the future research directions in the field of image segmentation techniques. Literature [32] discusses the positive role played by rough set theory in dealing with the complexity of the image segmentation process, and analyzes the current image segmentation methods with rough set theory as the core logic, pointing out that the image segmentation strategy based on rough set is more stable and has better performance. Literature [33] reviewed machine learning methods, such as Markov Random Fields, k-mean clustering, etc., and deep learning architectures in biomedical image segmentation techniques, analyzed the limitations and advantages of various strategies, and targeted heuristics to address the above limitations.

Some scholars have also explored the current status of design and practice of image segmentation techniques based on deep learning algorithms from the fields of image analysis, medical diagnosis, and agricultural pest monitoring. Literature [34] emphasizes the importance of image segmentation technology for geo-targeting-based image analysis (geoobia or OBIA), and also summarizes the existing segmentation tools and software packages, focuses on the optimal parameters and algorithm selection of image segmentation technology, and points out that the future optimization of segmentation algorithms is automated, accurate and efficient. Literature [35] conceived an image segmentation algorithm to realize automatic detection and classification of plant leaf diseases, which effectively reduces the work of agricultural disease detection and promotes the good development and production of agriculture. Literature [36] summarizes the research on the application of image segmentation technology in the field of cardiac medicine, in which the common urban and rural forms contain magnetic resonance imaging (MRI), computed tomography (CT), etc., focusing on the practice of image segmentation technology based on deep learning technology exists in the direction of the optimization to be optimized, which provides an important reference for the research of image segmentation technology in cardiac medicine diagnosis.

In this paper, based on the algebraic topology method and CNN-Transformer network, we realize the construction of an image segmentation network model based on asymmetric topology preservation and apply it to the field of medical image segmentation. The model adopts a parallel branching structure that applies CNN and Transformer, and the global dependencies and low-level spatial details can be effectively captured by utilizing shallow coding. Meanwhile, in order to fully integrate the global features and spatial detail information, a hybrid feature aggregation strategy (HFAS) is proposed, which mainly consists of an alternating attention module (AAM) based on spatial good-channel features and a topological boundary preserving module (TBPM). Finally, in order to verify the model

efficacy, this paper conducts comparison experiments and ablation experiments on the ATSNet model, and explores the optimization of the model performance in terms of both the number of network layers in the Transformer branch and the filtering method in the TBPM.

2. Image segmentation network model based on asymmetric topology preservation

In this paper, an asymmetric topology-preserving image segmentation network (ATSNet) model is proposed, where the network employs a parallel branching structure based on CNN and Transformer to efficiently capture global information and low-level spatial details in a shallower way. Meanwhile, in order to fully fuse the global features and spatial details, a hybrid feature aggregation strategy (HFAS) is proposed. In the hybrid feature aggregation strategy, the Alternate Attention Module (AAM) is proposed to focus on spatial and channel key features, and the Topological Boundary Preserving Module (TBPM) is constructed to explicitly extract boundary information.

2.1. Theory related to algebraic topology

Algebraic topology is a branch of mathematics that combines algebraic and topological concepts to study structural similarity. It is primarily concerned with how topological spaces can be inscribed and categorized by algebraic structures, and how topological concepts can be used to study algebraic structures.

2.1.1. Topological space. Topological space is a fundamental concept in mathematics used to study the structure and properties of spaces. In a topological space, the set of points and the sets of these points have certain structural relationships between them, and this relationship will be introduced into the open set to describe it. More complex topological spaces may be used in the study of networks or graphs in general, which has led to the use of topological spaces in several branches of mathematics, such as algebra, geometry, differential geometry, partial differential equations, and mathematical physics. In these fields, the study of topological spaces enables researchers to better understand and solve various practical problems. The definition of topological space is given below:

Definition 1: Consider a set of ordered pairs (X,τ), where X is a set and τ is a cluster of subsets of X, when the following properties are satisfied:

1) the empty set and X belongs to τ.

2) the sum of any number of elements in τ still belongs to τ.

3) the intersection of a finite number of elements in τ still belongs to τ.

Then this ordered pair (X,τ) is a topological space, τ is a topology on X, and the elements of τ are called open sets. Assuming $X=\{a,b,c\}$, then $\tau_1=\{\varnothing,\{a\},\{a,b,c\}\}$ is a valid topology on X, while $\tau_2=\{\varnothing,\{a\},\{b\},\{c\}\{a,b,c\}\}$ is not, since $\{a\}\cup\{b\}=\{a,b\}$ is not an element of τ. Then (X,τ_2) is a topological space. From this, it can be seen that only one topological space is constructed here, and in fact multiple topological spaces can be constructed depending on X and the different τ.

2.1.2. Simplexes and Simplex Complexes. A simplex is a multidimensional geometric shape, in n-dimensional space, a simplex is a convex polyhedron consisting of n vertices, each of which is a n-dimensional vector, which is mathematically represented as follows:

Definition 2: In a N-dimensional Euclidean space R^N, there is a set $\{v_0,v_1,...,v_n\}$ of points where $\forall t_i\in R$, $\sum_{i=0}^{n}t_i=0$ as well as $\sum_{i=0}^{n}t_iv_i=0$ hold if and only if $t_0=t_1=...=t_n=0$, then the points $v_0,v_1,...,v_n$ are geometrically independent. A n-dimensional simple complex σ consisting of them is a set satisfying the following conditions:

1) $x=\sum_{i=0}^{n} t_i v_i$, where $\sum_{i=0}^{n} t_i = 1$.

2) $\forall i, t_i \geq 0$, t_i is uniquely determined by x.

Based on the above definition, the shape of a low-dimensional simplex can be described intuitively; a 0-simplex is an independent point, a 1-simplex is a line segment joining two 0-simplexes, a 2-simplex is a triangle, and a 3-simplex is an orthotetrahedron, then analogously it can be introduced that a 4-simplex is made up of five points, and that it has a surface that is an orthotetrahedron.

Definition 3: In Euclidean space, there is a simple complex form M, a collection of simple forms, when the following conditions are satisfied:

1) Any face in M still belongs to M.

2) If the intersection of any two simplexes in M is not the empty set, then the intersection of these two simplexes is a face of both.

By the above definition, assuming that there is a simple complex M consisting of simplex $J=\{a,b\}$ and simplex $L=\{d,e,f\}$, then M can be denoted as $\{\{a\},\{b\},\{c\},\{d\},\{e\},\{f\},\{a,b\},\{a,c\},\{b,c\},\{c,d\}\{d,f\},\{d,e\},\{e,f\},\{a,b,c\},\{d,e,f\}\}$.

2.1.3. Other theories of algebraic topology:

1) Continuous mappings and congruences. Definition 4: Suppose that there are two topological spaces X and Y which satisfy the correspondence $f: X \to Y$, and f is regarded as a continuous mapping from X to Y if the original image of every open set of Y is an open set in X.

Definition 5: Suppose there are two topological spaces X and Y which satisfy correspondences $f: X \to Y$, f^{-1} is an inverse map of f, and f is a homoembryonic map if both f and f^{-1} are continuous, when $X=Y$, topological spaces X and Y are homoembryonic.

Topological spaces that are homogerms will have some special properties.

Definition 6: Suppose there are three topological spaces X, Y and Z which have the following properties:

(1) Constant homomorphisms: $X \to X$ is a homomorphism.

(2) If $X \to Y$ is a homomorphism, then its inverse mapping $Y \to X$ is also a homomorphism.

(3) If both $f: X \to Y$ and $g: Y \to Z$ are homomorphisms, then $g \circ X \to Z$ is also a homomorphism.

Definition 7: Suppose there are three topological spaces X, Y and Z which have the following properties:

(1) X and X are isomorphic.

(2) If X and Y are isomorphic, then Y and X are also isomorphic.

(3) If X and Y are isomorphic, and if Y and Z are isomorphic, then if X and Z are isomorphic.

2) Homology. Based on the definition of continuous mappings in the above section, the notion of homography can be derived:

Definition 8: Assume that there are two topological spaces X and Y, the existence of two continuous mappings $f: X \to Y$ and $g: X \to Y$, and a unit interval I, if there exists a continuous mapping $H: X \times I \to Y$ when the following conditions are satisfied:

(1) All $x \in X$'s have $H(x,0)=f(x)$ and $H(x,1)=g(x)$.

(2) H is a composite mapping of f and g on $X \times \{0\}$ and $X \times \{1\}$ respectively then f and g are homotopy mappings.

3) Continuous homotopy. From the concepts of simplex and simple complex, it can be noticed that there are some features of complex graphs such as holes and holes which are called topological features. In the study of this paper, these topological features are not static, it will appear or disappear

as the target changes. The persistent homotopy (PH) mentioned in this subsection is a method to find the number of target holes with a simple complex shape as the target.

In homology, homotopy groups are introduced to characterize these holes. Assume that there is a topological space X, and that the n-dimensional homotopy group $H_n(X)$ consists of all n-dimensional chains generated by mapping $(n+1)$ for the boundary of a simplex in X to a n-dimensional simplex. One of the important features of homotopy groups is that they are closed to group operations if a and b are elements in $H_n(x)$. Homotopy groups can be computed by chaining and boundary mapping with the formula:

$$H_n(x) = \ker(\partial : C_n(x) \to C_{n-1}(X)) / im(\partial : C_{n+1}(X) \to C_n(X)) \tag{1}$$

where $C_n(X)$ is the n-dimensional simplex complexity of X, and ∂ is the boundary mapping which maps n-dimensional chains to $(n-1)$-dimensional chains.

4) The Betti number. The Betti number is an important topological invariant in algebraic topology. In continuous cohomology, the Betti number is the rank of the cohomology group, which can be thought of as the number of dimensions of "holes" or "holes" in the space. For a n-dimensional topological space X, the ird Betti number $\beta_i(X)$ is defined as the rank of the homotopy group $H_i(X,Z)$ of X (i.e., the number of linearly independent generators in the homotopy group). It can take infinite or non-negative integers, where b_0 refers to the number of connectives, b_1 to the number of cavities in one dimension, and b_2 to the three-dimensional cavity, which is enclosed by surfaces and can also be interpreted as a "cavity". In addition, the value of b_0, b_1, b_2 changes as the value of parameter ε increases.

2.2. Transformer technology

Transformer technology is widely used in the field of natural language processing, which has become the mainstream model in this field, and it is also applicable to the field of computer vision, which has certain advantages over CNN architectures, such as the ability to solve the problem of long-distance dependence.

2.2.1. Self-attention mechanisms. The self-attention mechanism is one of Transformer's core techniques, which allows the model to dynamically weight and aggregate information from other locations as it processes inputs from each location. This mechanism allows the model to capture the dependencies between different locations in the input sequence, thus enabling the modeling of long-range dependencies.

The scaled dot product attention mechanism is an implementation of the self-attention mechanism. Before using the attention mechanism, the input image needs to be multiplied with the weight matrix to obtain the Q, K, and V matrices.Q, K, and V represent the query matrix, key matrix, and value matrix, respectively. Then, the attention score is calculated according to the rules, scaling as well as normalizing the score and finally multiplying it with the value matrix to get the final attention result.

The multi-head attention mechanism is an extended form of the self-attention mechanism, which introduces multiple parallel attention heads on top of the self-attention. First the input sequence is linearly transformed to obtain the query matrix Q, the key matrix K, and the value matrix V. Then the Q, K, and V matrices are partitioned separately into multiple heads, each of which has a dimension smaller than that of the original matrix. Next, the scaled dot product attention mechanism is used for each head individually to compute the attention output. Finally, the attention outputs of all heads are stitched together and linearly transformed to obtain the final multi-head attention output.

The multi-head attention mechanism enhances the expressive power of the model by applying multiple independent self-attention heads in parallel, each of which learns information about different aspects of the attention sequence. This parallel computation allows the model to capture the various

features and dependencies in the sequence more efficiently, and enhances the expressive power of the model by focusing on different features in the sequence from multiple perspectives at the same time.

2.2.2. Vision Transformer. Vision Transformer (ViT) model is a representative work of Transformer technology applied in computer vision [37].ViT model can effectively solve the problem of limited global context information capture ability and have better processing effect on larger scale images compared with traditional CNN architecture model.ViT model can easily deal with various sizes of input images without using convolutional layer or pooling layer as CNN to adapt to different sizes of image inputs. The ViT model can easily cope with input images of various sizes without the need to use convolutional or pooling layers to adapt to different sizes of image inputs as CNNs do. The model is mainly composed of three parts: Embedding layer, Transformer encoder and MLPHead.

1) Embedding layer: The Transformer processes a multidimensional sequence of vectors (Token sequence), which cannot directly process the complete image, so the complete image needs to be partitioned into a collection of image patches of appropriate size before passing it into the network, to obtain a vector sequence composed of image patches. The number of image patches depends on the segmentation size of the image patches, which is different in different networks. Class markers dedicated to classification are added to the vector sequence, and then positional embedding is added to obtain the final processed Token sequence.

2) Transformer Encoder: It consists of multiple iteratively stacked Transformer blocks, which learns the feature representation of the input image patch blocks and captures the dependencies and importance between different image blocks. It mainly consists of layer normalization, multi-head attention mechanism and feed forward neural network (FFN). Among them, FFN mainly consists of fully connected neural network and GELU activation function, which is also usually called multilayer perceptron (MLP).

3) MLP Head: It usually consists of multiple fully connected layers and activation functions that map previously learned feature representations to probability values for specific image classification. After the vector sequence is passed into the MLPHead, only the previously inserted class markers are extracted and manipulated to generate the final result.

2.2.3. Uniformer. Uniformer is a hybrid architecture that unifies convolutional operations and self-attention mechanisms, effectively combining the advantages of CNN and Transformer architectures, and can be used to process a wide range of image tasks [38].Uniformer proposes a new basic Block structure, which is mainly composed of a Dynamic Positional Encoder (DPE), MultiHeaded Relational Aggregator (MHRA), and feedforward neural network.

Traditional Transformer mainly relies on positional embedding to realize information encoding, however Uniformer usually uses dynamic position encoder to realize similar functions. The dynamic position encoder is mainly realized by a depth-separable convolution with an expansion value of 0. This is mainly due to the robustness of the depth-separable convolution to arbitrary inputs as well as its own lightweight nature. The computational flow of the input feature map in the dynamic position encoder is shown in Eqs. (2) to (3):

$$DPE(X_{in}) = DWConv(X_{in}) \tag{2}$$

$$X = DPE(X_{in}) + X_{in} \tag{3}$$

where X_{in} is the input feature map, DPE is the dynamic position encoder, and X is the output feature map after DPE processing.

Instead of using multi-head attention mechanism, Uniformer uses multi-head relation aggregator for feature processing, which can capture the global and local information in the target image more effectively. The multi-head relationship aggregator sets different tasks for different “heads” to pay attention to global and local information respectively, which enhances its multi-scale capturing ability

and improves the understanding of the image. The computation process of feature map in the multi-head relationship aggregator is shown in Equation (4):

$$Y = MHRA(Norm(X)) + X \tag{4}$$

where MHRA is the multi-head relationship aggregator, Norm is the normalization operation, and Y is the feature map output after being processed by MHRA.

The feedforward neural network is consistent with the traditional ViT model, which is composed of two Linear layers and a GELU activation function. The computational flow of the feature map in the feedforward neural network is shown in Equation (5):

$$Z = FFN(Norm(Y)) + Y \tag{5}$$

where FFN is feedforward neural network and Z is the feature map output after processing by FFN.

2.2.4. Swin Transformer. Swin Transformer is a novel architecture for visual Transformer. Different from the traditional visual Transformer, the sliding window mechanism is introduced in Swin Transformer to overcome the high computational complexity problem caused by global self-attention computation [39]. The core innovation of the sliding window mechanism is to divide the image into fixed-size non-overlapping windows, perform self-attention computation within each window, and at the same time realize cross-window information interaction by sliding the position of the window between different layers. This approach not only significantly reduces the computational effort, but also retains the advantage of capturing long-range dependencies in the original model.Each module in Swin Transformer consists of two subunits, each containing four components: layer normalization, attention computation, layer normalization, and a multilayer perceptron layer. The first subunit uses the Window Multihead Self-Attention (W-MSA) module, while the second subunit employs the Shift Window Multihead Self-Attention (SW-MSA) module. This design effectively achieves efficient extraction and fusion of local and global features by alternating between window and shift window attention mechanisms in different subunits.

2.3. ATSNet based image segmentation method

2.3.1. Network construction. The overall structure of the ATSNet network is shown in Figure 1. In the encoder part, the information is processed differently using a hybrid network combining parallel branches CNN and Transformer.The CNN branch enlarges the receptive field by successive downsampling and encodes the features from local to global. In the Transformer branch, the feature mapping is serialized and then the input is reset into a series of flat blocks, positionally encoding the block embeddings, which are then fed into the Transformer, recovering local details while modeling global dependencies. Subsequently, features from different branches with the same resolution are fed into the proposed Hybrid Feature Aggregation Strategy (HFAS) for joint prediction, which consists of an Alternate Attention Module (AAM) and a Topological Boundary Preserving Module (TBPM). Finally, in the decoder section to avoid the loss of detailed features, the resolution of the original input feature map is gradually restored using cascaded up-sampling and the segmentation map is generated using gated jump connections in combination with a multilayer feature map.

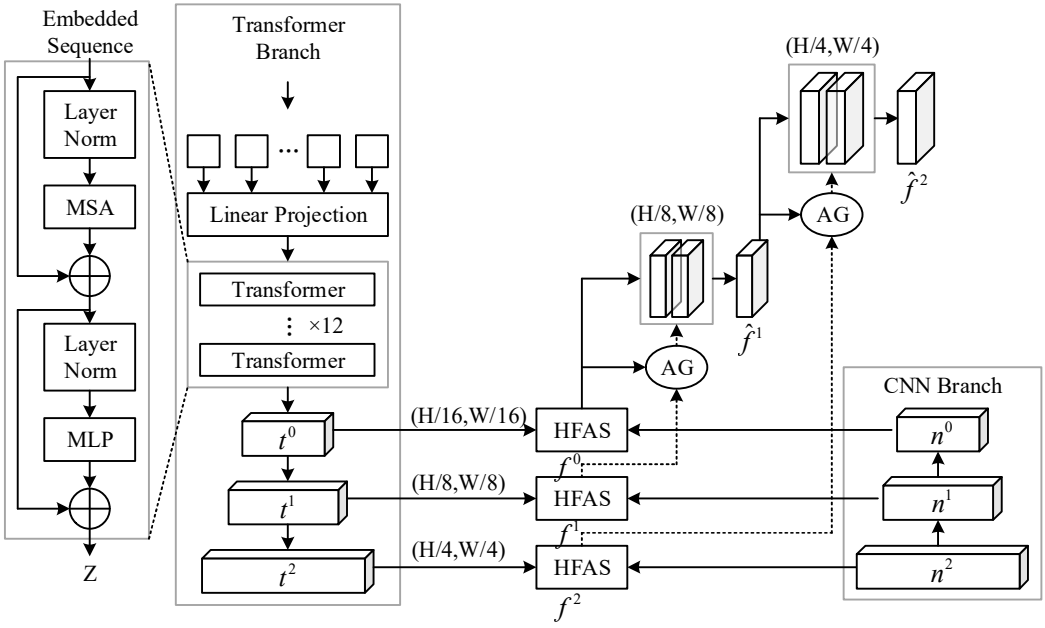


Fig. 1. Overall network structure

2.3.2. Transformer branch. The Transformer branch continues the typical encoder-decoder paradigm, which treats each pixel of an image as an element in a sequence and constructs the sequence based on the relationship between its neighboring positions. In this paper, the feature mapping is serialized, and then the input feature map X is reset into a series of flat blocks, each of size $P \times P$, where P is typically set to 16, indicating the length of the input sequence. The blocks are then flattened and mapped to the underlying D-dimensional embedding space by linear projection. At the same time, the block embeddings are encoded and the following positional information is preserved:

$$Z_0 = [X^1 E, X^2 E, ..., X^N E] + E_{pos} \tag{6}$$

where $E \in R^{(P^2 C)} \times D$ represents the block embedding projection and $E_{pos} \in R^{N \times D}$ represents the positional embedding.

The self-attention mechanism is one of the core components of Transformer, which can be used to order the relationships between different elements in the model. The self-attention formula is expressed as follows:

$$Attention(Q, K, V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

For the decoder part, cascaded up-sampling is used to recover the spatial resolution of the input feature maps,, in the Transformer branch, the feature maps 1 and 2 are processed to obtain the feature maps $t^1 \in R^{(\frac{H}{8}, \frac{W}{8} \times D_1)}$ and $t^2 \in R^{\frac{H}{4} \times \frac{W}{4} \times D_2}$ respectively at the output side retaining the feature maps t^0, t^1, t^2 at different scales, and the CNN branch is fused sequentially corresponding to the output feature map n^0, n^1, n^2.

2.3.3. CNN branching. Generally CNN-based networks acquire a larger sensory field and extract global contextual information by continuously deepening the network, which leads to redundant computations and also loses some spatial details [40]. Considering that Transformer has long-range modeling capabilities, it fully benefits from the global self-attention mechanism included in the model. Therefore, in the network of this paper, ResNet-34 is used as the backbone of the CNN branch and the

outputs of the first 4 layers are retained. The output feature maps of blocks $[(n^0 \in R^{\frac{H}{16}\times\frac{W}{16}\times C0})$, $3(n^1 \in R^{\frac{H}{8}\times\frac{W}{8}\times C1})$, and $2(n^2 \in R^{\frac{H}{4}\times\frac{W}{4}\times C2})$ are fused with the results of the corresponding Transformer layer, allowing the network to utilize the Transformer branch to obtain global contextual information. The design achieves the effect of killing two birds with one stone, without losing the ability to localize the underlying details and improving the efficiency of global context modeling.

2.3.4. Hybrid Feature Aggregation Strategy. In order to effectively fuse the coded information of CNN and Transformer, this paper proposes a new hybrid feature aggregation strategy (HFAS), the specific structure of which is shown in Fig. 2(a), which consists of two sub-modules, namely: the Alternate Attention Module (AAM), which focuses on the important information, and the Topological Boundary Preserving Module (TBPM), which explicitly learns the edge features and preserves the topological structures The two modules work together on the CNN branch and the Transformer branch.Hadamard operations enable more detailed information extraction and feature learning through the interaction between the features of the two branches. Different features are integrated through interactive modeling to improve the performance and effectiveness of the model, denoted as $\hat{m}^i$. Finally, the three branches are connected through the residual module and passed to the fusion feature f^i, which operates as follows:

$$\hat{t}^i = AAM(t^i), \hat{n}^i = AAM(n^i) \tag{8}$$

$$\hat{m}^i = Conv(t^i W_1^i \otimes n^i W_2^i), f^i = \mathrm{Re}sidual([\hat{m}^i, \hat{t}^i, \hat{n}^i]) \tag{9}$$

In the above equation, $W_i^i \in R^{Di\times Li}, W_2^i \in R^{Ci\times Li}$, represents the two-branch input matrix, Conv denotes the 3×3 convolutional layer, and "$\otimes$" represents the Hadamard operation.

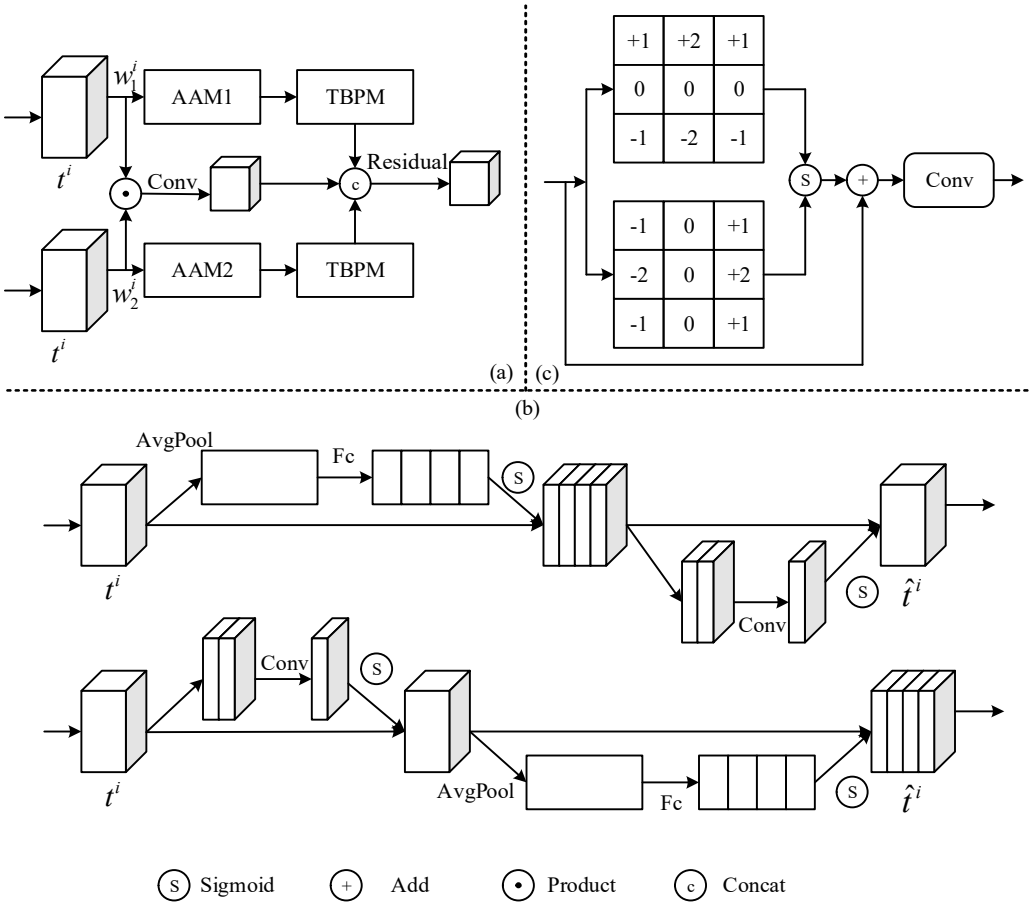


Fig. 2. Structure of mixed feature aggregation strategy

1) Alternating Attention Module. As shown in Fig. 2(b), the Alternate Attention Module (AAM) uses spatial attention and channel attention alternately on the CNN and Transformer branches. In the Transformer branch, the channel attention mechanism is first used to improve the global information of the Transformer branch, and then the spatial attention mechanism is used to enhance the feature representations at different spatial locations In the CNN branch, switching the order of the modules, the spatial attention is first used to suppress irrelevant regions because the low-level features of the CNN may be noisy, and then the channel attention to enable the network to focus more accurately on the channels that are most valuable for the task at hand, improving the network's representation. It is worth noting that the channel attention implementation uses SE blocks and the spatial attention is taken from CBAM blocks as spatial filters. In this paper, input features t^i and n^i are represented as $\hat{t}^i$ and $\hat{n}^i$, respectively, after being processed by the alternating attention block.

2) Topological Boundary Preserving Module. In order to further improve the boundary feature extraction capability, this paper introduces the Topological Boundary Preserving Module (TBPM) as shown in Fig. 2(c). The double-branch features processed by AAM are input into TBPM to filter out the trivial information unrelated to the boundary and extract the necessary boundary features. The horizontal and vertical edge features in the image are detected using the Sobel operator, and the gradient maps M_x and M_y are obtained for the horizontal G_x-direction and the vertical G_y-direction, respectively.The Sobel operator is a commonly used and relatively simple boundary detection operator, which detects the gradient changes in the image by detecting the gradient changes in the image and converting them into the gradient intensity and direction of each pixel in the image. The two convolutions are defined as:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{10}$$

The image normalized by the gradient mapping is processed with the Sigmoid function and then the input feature map is fused with it to obtain the boundary enhanced feature map $\hat{F}_e$:

$$\hat{F}_e = F_e \otimes \sigma(M_{xy}) \tag{11}$$

where ' $\otimes$ ' denotes the Hadamard operation, σ denotes the Sigmoid function, and gradient mappings M_x and M_y cascade along the channel dimension to obtain M_{xy}.

2.3.5. Loss function. Since different segmentation tasks use different loss functions, this paper defines a total loss function to jointly optimize the segmentation loss and boundary loss of an image.

1) Segmentation loss. According to different target segmentation, a weighted sum of binary cross-entropy loss L_{bce} and cross-parity ratio loss L_{loU} is used. It is defined as:

$$L_{bce} = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{12}$$

$$L_{IOU} = 1 - \frac{(y_i \times \hat{y}_i)}{(y_i + \hat{y}_i - y_i \times \hat{y}_i)} \tag{13}$$

$$L_{seg} = L_{lOU}^w + L_{bce}^w \tag{14}$$

where y_i and $\hat{y}_i$ denote the true value and predicted label of the ird pixel, respectively. L_{lOU}^w and L_{kc}^w denote the weighted IoU loss and the binary cross entropy (BCE) loss.

2) Boundary loss. The Dice coefficient is a good measure of the accuracy of the model boundary prediction, especially for image segmentation tasks where fine boundaries need to be precisely localized. In addition, Dice loss can also solve the class imbalance problem due to foreground and background when calculating the boundary, which effectively improves the accuracy and robustness of the model. Therefore, Dice loss is introduced in this paper to further optimize the segmentation task. It is defined as follows:

$$L_{Dice} = 1 - \frac{2(y_i \times \hat{y}_i)}{(y_i + \hat{y}_i)} \tag{15}$$

Same as above equation, y_i and $\hat{y}_i$ denote the true value and predicted label of the ird pixel respectively.

3) Total loss. Segmentation loss L_{seg} and boundary loss L_{Dice} constitute the total loss. The prediction results are obtained by applying the depth monitoring strategy in the first fusion branch and the Transformer branch, respectively. In summary, the total loss is defined as follows:

$$L = \alpha L_{seg}(G, head(\hat{f}^2)) + \gamma L_{seg}(G, head(t^2)) + \beta L_{seg}(G, head(f^0)) + L_{Diee} \tag{16}$$

where α, γ, β are all adjustable parameters and G represents the truth value.

3. Model application analysis

In order to verify the effectiveness of the proposed CNN-Transformer based asymmetric topology preserving segmentation network (ATSNet) model, this paper takes medical images as the object to conduct image segmentation experiments, which specifically include comparison experiments with other models, clinical application experiments, network structure ablation experiments, network effectiveness verification, and comparison experiments of different filtering methods.

3.1. Comparative Experiments and Analysis

In this paper, three public datasets, LiTS2017, BraTS2020, and ACDC, and a private dataset consisting of MR1 scans of endometrial cancer provided by a hospital are used for comparison experiments.

Among them, the LiTS2017 dataset is a public dataset for liver tumor segmentation containing 150 3D CT scan images and corresponding liver and tumor segmentation masks, each CT scan usually consists of 42 to 1026 slices.The BraTS2020 dataset is a public brain tumor segmentation dataset containing 412 MRI scans per case including MRI images in four modalities, T1, T1 enhancement, T2, and FLAIR, and the segmentation mask needs to differentiate between enhanced tumor (ET), tumor core (TC), and whole tumor (WT).The ACDC dataset consists of MRI cardiac images from 200 patients, which include the four chambers of the heart and myocardial tissues labeled into three categories: right ventricular cavity (RV), left ventricular cavity (LV) and myocardium (Myo). The endometrial cancer dataset contains MR1 scans of 81 pathologically diagnosed cases of endometrial cancer with TICE and T2-weighted modalities.

In this paper, we take medical images as an example for image segmentation experiments, and use Dice and HD95 (95% Hausdorff Distance) to evaluate the segmentation performance of medical images.Dice is derived by calculating the overlap area between the model prediction result and the real segmentation result than the total area of the two, and its value ranges from 0 to 1, and the larger the value means that the segmentation result is more accurate.HD95 is the measure of the segmentation result and the real segmentation result, and it is the measure of the segmentation result and the real segmentation result. is a measure of the distance between the segmentation result and the real segmentation result, which is calculated by finding the maximum value of the distance

between the 95% farthest point in the segmentation result and the nearest point in the real segmentation result, and the smaller the value is, the more accurate the segmentation result is.

The segmentation performance comparison results of different methods on the ACDC dataset are shown in Table 1.

As can be seen from Table 1, the ATSNet method used in this paper performs the best, with an average Dice value of 93.03%. Compared with the UNet method, the average Dice coefficient of the ATSNet method is improved by 5.68%. This indicates that the method proposed in this paper can obtain more accurate segmentation results on the ACDC dataset.

Table 1. Comparison of segmentation performance of different methods on ACDC dataset

Methods	Mean Dice (%) ↑	Dice (%) ↑		
		RV	Myo	LV
U-Net	87.35	79.83	95.24	86.97
Att-UNet	87.44	79.84	94.80	87.69
VIT	87.79	82.17	94.59	86.62
TransUnet	89.85	84.62	96.21	88.71
SwinUnet	89.80	84.53	96.77	88.10
Missformer	88.63	86.77	92.45	86.68
C^2former	91.78	89.71	95.51	90.12
TT-Net	92.40	90.97	95.26	90.98
ATSNet	93.03	91.32	95.57	92.19

The results of quantitative comparison of this paper's algorithm with other state-of-the-art methods on LiTS2017 dataset are shown in Table 2, and the algorithms in this chapter have been consistently improved in terms of accurate segmentation. By comparing the Dice coefficient and HD, this paper's method achieves the best performance in liver and tumor segmentation. In short, the algorithm proposed in this chapter outperforms other state-of-the-art methods on the LiTS2017 dataset.

Table 2. Comparison of the different methods on the LiTS2017 data set

Methods	Dice (%) ↑		HD (mm) ↓	
	Liver	Tumor	Liver	Tumor
U-Net	76.98	66.99	11.22	19.12
TransUnet	89.77	66.65	6.41	18.28
SwinUnet	86.64	63.04	7.93	25.67
VIT	90.42	67.33	6.77	19.30
ATSNet	93.54	68.56	5.43	14.65

And the results of the comparison between the ATSNet algorithm proposed in this paper and other state-of-the-art methods on the BraTS2020 dataset are shown in Table 3. The results show that this paper's method achieves better performance compared with other existing methods, with the Dice values of 81.51%, 90.53% and 83.19% for ET, ET and TC sites, respectively, which are larger than those of the other algorithms, while the corresponding HD values of 16.08 mm, 5.15 mm and 8.73 mm, respectively, which are smaller than those of the other algorithms. This proves its superior performance in multimodal MRI data segmentation tasks, which can help doctors diagnose and treat brain tumors and other related diseases more accurately.

Table 3. Comparison of the different methods on the BraTS2020 data set

Methods	Dice (%) ↑			HD (mm) ↓		
	ET	WT	TC	ET	WT	TC
3D U-Net	69.12	83.32	79.34	51.09	13.21	13.45
Basic V-Net	61.68	84.38	76.48	47.87	20.92	11.82
Deeper V-Net	68.55	86.17	77.62	44.92	13.45	16.2
Residual 3D V-Net	71.82	82.25	76.88	36.37	13.04	13.28
Transbts	77.92	88.59	81.56	17.15	6.48	10.63
ATSNet	81.51	90.53	83.19	16.08	5.15	8.73

The results of the comparison between the algorithm proposed in this paper and other methods on endometrial cancer dataset are shown in Table 4. The results show that compared to other methods, the algorithm in this chapter achieves better segmentation results in terms of evaluation metrics such as Dice coefficient and Hausdorff distance. In particular, the algorithm in this paper improves the segmentation performance of the uterus by 14.43% compared to the TransUNet method, and these results indicate that the algorithm in this chapter has a better performance in the task of endometrial cancer segmentation, which can help doctors to more accurately diagnose and treat endometrial cancer and other related diseases.

Table 4. Comparison of the different methods on the endometrial carcinoma data set

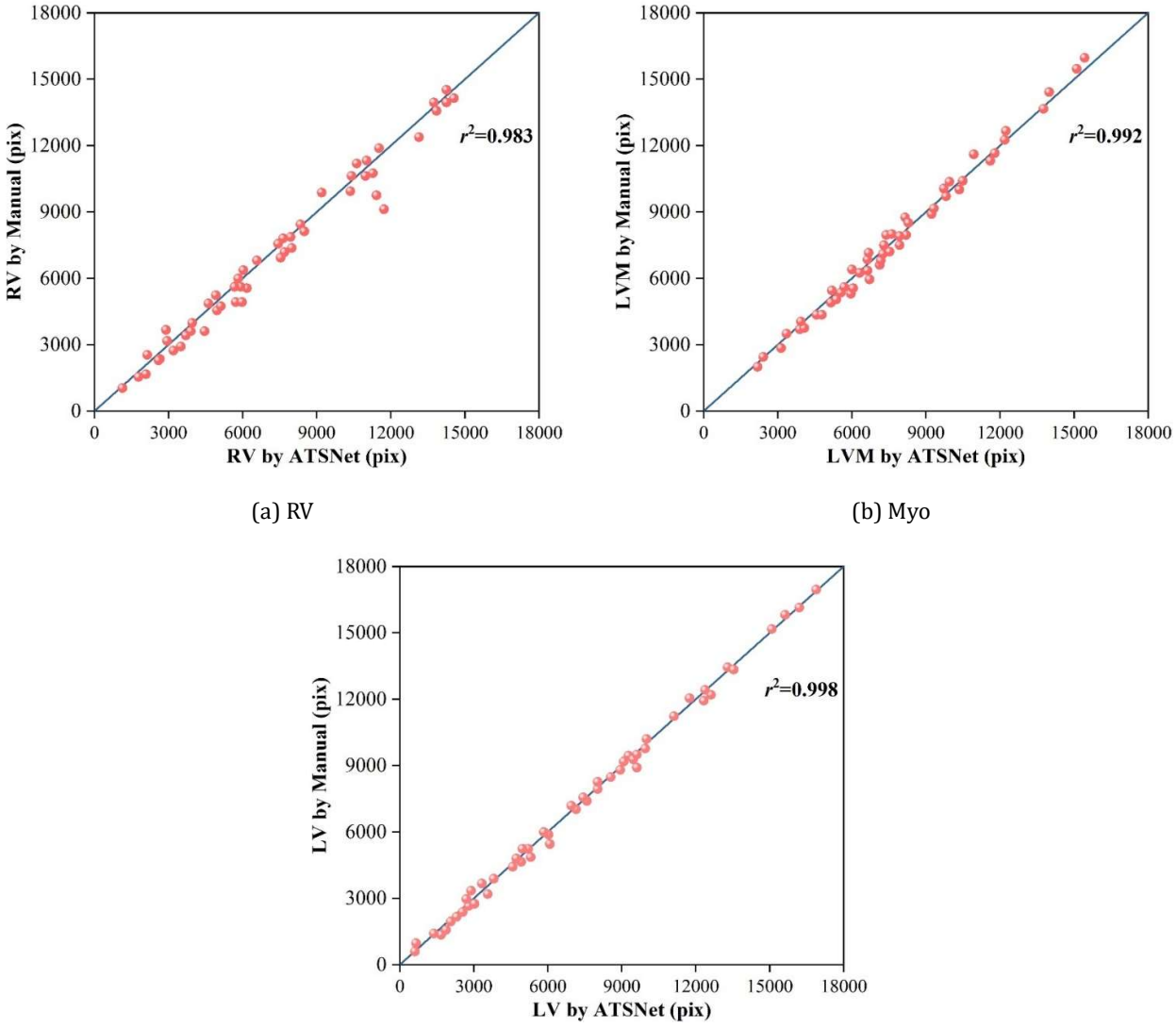
Methods	Dice (%) ↑		HD (mm) ↓	
	Uterus	Tumor	Uterus	Tumor
U-Net	71.61	85.06	10.16	7.15
TransUnet	64.54	86.71	9.87	9.86
SwinUnet	67.99	85.73	16.46	8.95
VIT	67.22	84.21	12.75	8.43
ATSNet	78.97	90.13	7.35	6.14

3.2. Analysis of clinical applications

In this section, to assess whether the ATSNet algorithm has the capability to be used as a potential method in clinical diagnosis, a quantitative analysis experiment was conducted to evaluate the segmentation results of 50 samples using the Bland-Altman method and correlation coefficients and to compare them with the results of manual annotation by clinicians.

The study conducted a quantitative analysis experiment using the ACDC dataset, and the results of the comparative analysis of the ATSNet algorithm with the clinically annotated correlation coefficients are shown in Figure 3. Where, (a)~(c) represent the results of the analysis of the three indicators of RV, Myo and LV, respectively. The closer the correlation coefficient r is to 1 indicates a stronger correlation.Bland-Altman plots were used to evaluate the differences between the two methods and to calculate the upper and lower limits of the mean difference versus 95%.

The RV, Myo and LV measurements from the algorithm of this paper and the manual labeling were compared and found to be highly correlated. Comparing and Manual's measurements of RV, Myo, and LV (N=50), the ATSNet and Manual measurements of all three metrics were highly correlated, with r2 of 0.983, 0.992, and 0.998 for RV, Myo, and LV, respectively.



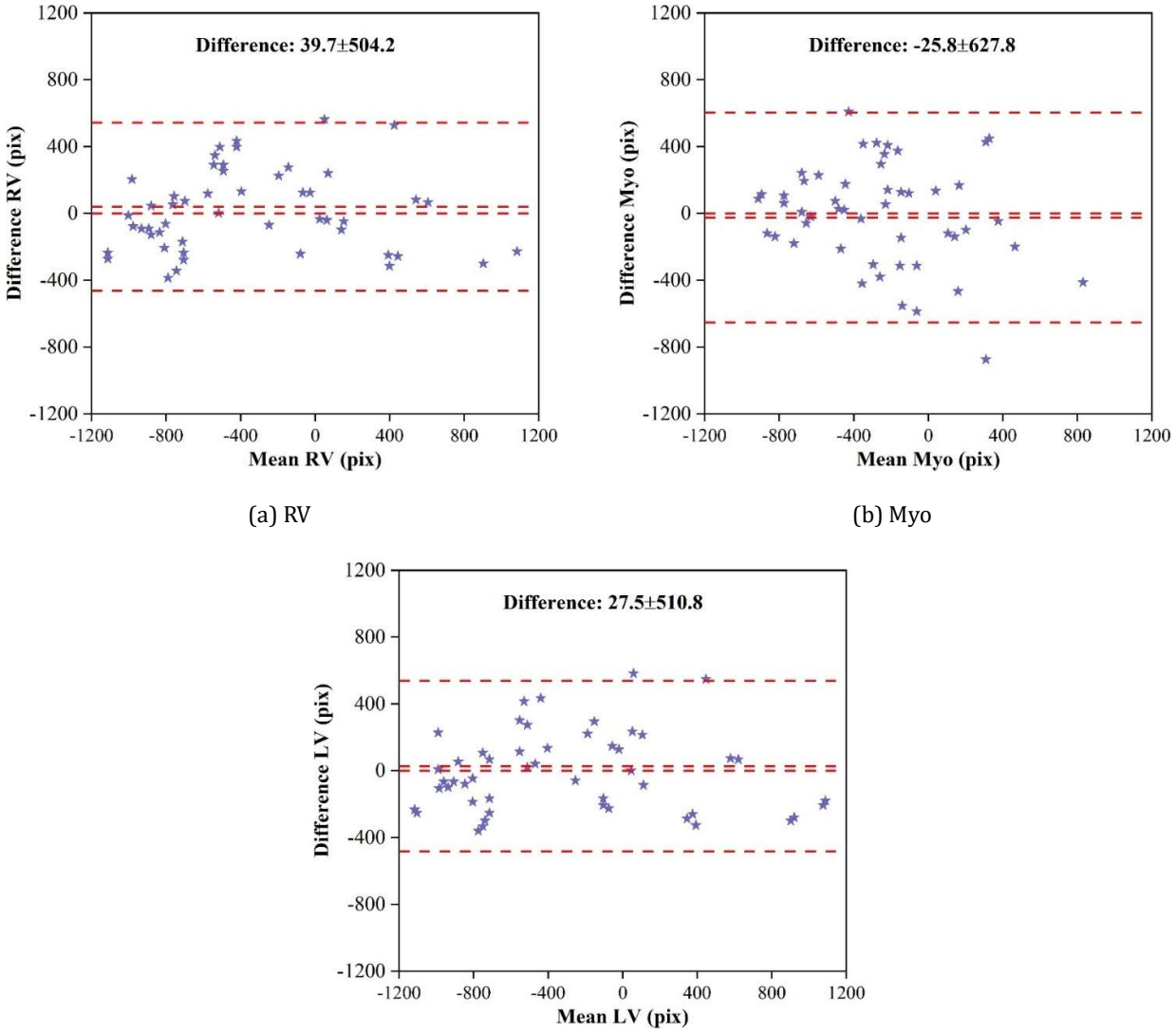
(a) RV (b) Myo

(c) LV

Fig. 3. Analysis of correlation coefficient between ATSNet and clinical annotation

The results of RV and Myo by paired samples t-test showed that there was no statistically significant difference between RV indicators and Myo indicators in terms of ATSNet and Manual measurements (T_{RV}=-0.457, P_{RV}=0.664>0.05, T_{Myo}=-0.573, P_{Myo}=0.582>0.05).The results of LV indicators by Wilcoxon signed-rank test results showed that there was no statistical difference between ATSNet and Manual measurements of LV indicators (Z=-0.516, P=0.627>0.05). In conclusion, there was no statistical difference between the ATSNet and Manual measurements of the three indicators, which verified the validity of the methodology of this paper.

In addition, the consistency results between this paper's algorithm and clinical experts were analyzed using Bland-Altman as shown in Figure 4. (a)~(c) represent the agreement results for the three metrics of RV, Myo and LV, respectively. Figure 4 calculates the 95% limits of agreement (95% LoA) using the Bland-Altman method, i.e., 95% of the variance is considered insignificant in clinical applications. The results showed that the 95% limits of agreement between ATSNet and Manual measurements using the Bland-Altman method ranged from -464.5pix to 543.9pix for RV, -653.6pix to 602.0pix for Myo, and -483.3pix to 538.3pix for LV, and that in at least 48 of 50 samples, the samples were within 95% LoA, indicating that the segmentation results of the algorithm in this paper can be substituted for clinicians' manual labeling and can be accepted in clinical practice.



(a) RV

(b) Myo

(c) LV

Fig. 4. Bland-Altman analysis between ATSNet and the clinical experts

3.3. Network structure ablation experiments

In order to validate the effectiveness of each part of the ATSNet network structure, the intravascular ultrasound (IVUS) image segmentation was used as an experimental object to perform ablation experiments on the network structure, which were conducted on the backbone network including only CNN encoders, the backbone network including only Transformer branches, the backbone network including fusion branches of CNNs and Transformers, and the models combining CNN-Transformer parallel branching networks with hybrid feature aggregation strategies are experimented.

The ablation experiments were evaluated on a dataset consisting of 852 IVUS data for training and testing. All images were acquired by a commercially available imaging system with a raw image resolution size of 500 pixel × 500 pixel.The standard lumen and plaque regions of the dataset were manually labeled by two researchers under the guidance of a physician specializing in cardiovascular disease and divided into a 3:1 ratio between the training set and the test set.

Meanwhile, in order to evaluate the performance of the algorithm, based on the $Dice$ coefficient and Hausdorff distance (HD) used in the previous comparison experiments, three evaluation metrics are introduced to evaluate the model segmentation results, such as the Jaccard metric (JM), the

percentage of the area difference (PAD), and the topological structure error rate (TER). $Dice$ and JM mainly measure the area difference between the segmentation result and the true value, the values of both coefficients are in the range of 0 to 1, with larger values indicating higher accuracy.HD mainly measures the difference between the contour of the segmentation result and the contour of the true value.PAD measures the area difference between the segmentation result and the label of the true value.RTER is a measure of the segmentation result's compliance with the a priori knowledge of the medical sciences.The values of HD, PAD, and The smaller the value of TER, the closer the segmentation result is to the true value label.

The results of the ablation experiments are shown in Table 5. Where Lum denotes the segmentation results of the inner membrane and Med denotes the segmentation results of the middle membrane.

As can be seen from Table 5, the backbone network based on a pure CNN encoder can obtain more satisfactory segmentation results, and the maximum values of the Dice and JM metrics reach 0.936 and 0.902, respectively.However, the CNN also has some limitations, i.e., for the targets in the complex background, it may not be able to accurately capture the long-distance dependence relationship between the different locations, which restricts the further improvement of the model accuracy , and thus its HD, PAD, and TER values are relatively large. The backbone network of pure Transformer encoder, on the other hand, lacks the local feature extraction capability of CNN, which makes it difficult to capture local spatial localization and texture information in the image, and often needs to be trained on large-scale data, and thus the segmentation effect is not satisfactory. Therefore, in the actual image segmentation task, CNN and Transformer have their respective advantages and shortcomings, and the combination of Transformer model and CNN can comprehensively utilize the advantages of both. The backbone network containing fusion branches of CNN and Transformer provides a priori knowledge through a unified coding layer to overcome the lack of a priori knowledge when Transformer does not have large-scale pre-training, and incorporates the results of the same CNN segmentation into each Transformer layer, so as to combine the detailed features output from the CNN coding branch with the detailed features output from the Transformer coding branch and the detailed features output from the CNN coding branch. Transformer encoding branch outputs of global features are further fused. It can be seen that this proposed backbone network better combines the advantages of CNN and Transformer, resulting in further improvement of model accuracy.

Although the introduction of global attention can largely improve the accuracy of segmentation, however, there are still cases of mis-segmentation without topological constraints. In the overall test images obtained from the experiments using only the backbone network, about 2.8% of the images still have blurred edges or have mis-segmented target blobs in the background. Although the area of these mis-segmented blobs is relatively small, there is still the possibility of generating segmentation results that do not conform to the a priori knowledge of medicine, which will have an impact on the clinical parameter computation and visualization effects. The introduced topological hybrid feature aggregation strategy can well solve this problem, on the basis of using CNN and Transformer dual-branch parallelism, improving the global information of the Transformer branch and enhancing the feature representation of different spatial locations through the Alternate Attention Module AAM, and improving the boundary feature extraction ability by utilizing the Topological Boundary Preserving Module TBPM, so that the overall The segmentation accuracy is further improved, and without any post-processing operation, not only the occurrence of multi-region mis-segmentation is avoided, but also the edges of the segmented endothelium and midthelium are smoother and more fluent, which are more in line with the actual morphology of the blood vessels, thus further improving the visual effect.

Table 5. Results of ablation experiment

Model	Dice ↑		JM ↑		HD ↓		PAD ↓		TER ↓
	Lum	Med	Lum	Med	Lum	Med	Lum	Med	All
CNN	0.935	0.936	0.882	0.902	0.239	0.352	0.098	0.091	8.5%
Transformer	0.932	0.921	0.871	0.854	0.211	0.341	0.112	0.161	9.1%
CNN+Transformer	0.940	0.962	0.886	0.916	0.106	0.129	0.069	0.082	2.8%
ATSNet	0.944	0.971	0.897	0.919	0.101	0.102	0.064	0.074	0

3.4. Comparison of different network layers in Transformer branches

In the experiments, the total number of layers of the Transformer branching network is empirically kept as 5. However, when the number of Transformer layers in the network is 5, the problem of not being able to train due to the large amount of computation required may occur due to the lack of sufficient device memory. The segmentation results for different number of Transformer layers are shown in Table 6. It can be seen that when the number of Transformer layers is 1 or 2, although the overall accuracy is improved compared to the network with 0 Transformer layers, the improvement is not as significant as when the number of layers is 3 due to the introduction of less global information. However, when the number of layers is 4, the achievable accuracy is also not as good as the accuracy when the number of layers is 3, so the branch network with 3 Transformer layers is selected. It is hypothesized that there may be less data in the training set, the model is too complex resulting in overfitting, the efficiency of model training is reduced, and the structure with more Transformer layers instead fails to obtain better results.

Table 6. Segmentation results of different Transformer layers

Transformer layers	Dice ↑		JM ↑		HD ↓		PAD ↓	
	Lum	Med	Lum	Med	Lum	Med	Lum	Med
0	0.937	0.948	0.881	0.902	0.243	0.367	0.088	0.096
1	0.939	0.949	0.883	0.902	0.222	0.411	0.090	0.085
2	0.939	0.952	0.885	0.911	0.171	0.225	0.072	0.096
3	0.942	0.954	0.889	0.918	0.112	0.14	0.069	0.079
4	0.939	0.950	0.882	0.894	0.168	0.188	0.068	0.091

3.5. Comparison of Different Filtering Methods in Topological Boundary Preservation Module

In order to compare the effects of different filtering methods on the segmentation results, experiments are conducted using Gaussian filter layers and bilateral filter layers respectively in the topological boundary maintenance module, and the comparison results of different filtering methods are shown in Table 7. It is worth noting that the weights of the filter layers need to be frozen during the training process, no matter which filtering method, if they are involved in the updating of the gradient, these filter layers are equivalent to a convolutional layer loaded with initial weights, and the final result is not as good as the filter layer used only for smoothing the image. As can be seen from Table 7, the bilateral filter layer in the image segmentation of the inner membrane, corresponding to the value of Dice, JM, HD, and PAD metrics are 0.946, 0.893, 0.108, and 0.075, respectively, which are better than the image segmentation results of the Gaussian filter layer. The segmentation results of bilateral filter layer are also better than Gaussian filter layer in the image segmentation of the mesopelagic membrane. This is because the bilateral filter can better preserve the edges while smoothing the image, so more accurate results can be harvested when sampling the differential homogeneous deformation layer, which further improves the final segmentation results.

Table 7. Comparison of different filtering modes in TBPM module

Model	Dice ↑	JM ↑	HD ↓	PAD ↓

	Lum	Med	Lum	Med	Lum	Med	Lum	Med
Gaussian smoothing	0.951	0.948	0.891	0.907	0.112	0.121	0.082	0.086
Bilateral smoothing	0.946	0.955	0.893	0.915	0.108	0.109	0.075	0.081

4. Conclusion

In this paper, the algebraic topology method, CNN and Transformer algorithms are combined to construct ATSNet, an asymmetric topology-holding image segmentation network model with CNN and Transformer branches in parallel, and a hybrid feature aggregation strategy (HFAS) is proposed to achieve the improvement of image segmentation accuracy.

Comparing with other models, the ATSNet model used in this paper achieves optimal image segmentation results on four datasets, including LiTS2017, BraTS2020, ACDC, and endometrial cancer dataset, as well as the maximum value of Dice coefficient and the minimum value of HD value. In the clinical application, the correlation coefficients of this algorithm and the manually labeled RV, Myo and LV measurements reached 0.983, 0.992, 0.998, respectively, and there was no statistically significant difference between this algorithm and the manually labeled results in the three indexes of RV, Myo, and LV ($P>0.05$), which indicates that the image segmentation results of this algorithm are very close to the clinician's manually labeled results, and can be applied in clinical practice. It can be used in clinical practice.

At the same time, the model ablation experiments show that the branching parallel method combining CNN and Transformer in this paper is highly practical, and the proposed hybrid feature aggregation strategy can realize the full fusion of global features and spatial details, and give full play to the respective advantages of CNN and Transformer, which can greatly improve the image segmentation accuracy.

In addition, in the Transformer branch, the overall accuracy is improved the most when the number of layers of the Transformer network is three. In the topological boundary preservation module, the use of bilateral filtering can better preserve the edges and improve the image segmentation effect.

Funding

1. This work was supported by Shanghai Urban Construction Vocational College's Strengthening Teachers Program for Vocational Bachelor's Degree (AA-05-2024-02-17-01);

2. This work was supported by Shanghai Urban Construction Vocational College 2024 Annual School-level Research Projects (AA-02-2024-01-16-47).

References

[1] Tian, H., Wang, T., Liu, Y., Qiao, X., & Li, Y. (2020). Computer vision technology in agricultural automation—A review. *Information Processing in Agriculture*, 7(1), 1-19.

[2] Shi, X., Tang, K., & Lu, H. (2021). Smart library book sorting application with intelligence computer vision technology. *Library Hi Tech*, 39(1), 220-232.

[3] Hassaballah, M., & Hosny, K. M. (2019). Recent advances in computer vision. *Studies in computational intelligence*, 804, 1-84.

[4] Chai, J., Zeng, H., Li, A., & Ngai, E. W. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6, 100134.

[5] Xu, S., Wang, J., Shou, W., Ngo, T., Sadick, A. M., & Wang, X. (2021). Computer vision techniques in construction: a critical review. *Archives of Computational Methods in Engineering*, 28, 3383-3397.

[6] Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1), 7068349.

[7] Kakani, V., Nguyen, V. H., Kumar, B. P., Kim, H., & Pasupuleti, V. R. (2020). A critical review on computer vision and artificial intelligence in food industry. *Journal of Agriculture and Food Research*, 2, 100033.

[8] He, W., Jiang, Z., Ming, W., Zhang, G., Yuan, J., & Yin, L. (2021). A critical review for machining positioning based on computer vision. *Measurement*, 184, 109973.

[9] Szeliski, R. (2022). Computer vision: algorithms and applications. Springer Nature.

[10] Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., ... & Socher, R. (2021). Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1), 5.

[11] Khan, A. A., Laghari, A. A., & Awan, S. A. (2021). Machine learning in computer vision: a review. *EAI Endorsed Transactions on Scalable Information Systems*, 8(32), e4-e4.

[12] O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., ... & Walsh, J. (2020). Deep learning vs. traditional computer vision. *In Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1* (pp. 128-144). Springer International Publishing.

[13] Dhanya, V. G., Subeesh, A., Kushwaha, N. L., Vishwakarma, D. K., Kumar, T. N., Ritika, G., & Singh, A. N. (2022). Deep learning based computer vision approaches for smart agricultural applications. *Artificial Intelligence in Agriculture*, 6, 211-229.

[14] Dong, C. Z., & Catbas, F. N. (2021). A review of computer vision–based structural health monitoring at local and global levels. *Structural Health Monitoring*, 20(2), 692-743.

[15] Brunetti, A., Buongiorno, D., Trotta, G. F., & Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300, 17-33.

[16] Gite, S., Mishra, A., & Kotecha, K. (2023). Enhanced lung image segmentation using deep learning. *Neural Computing and Applications*, 35(31), 22839-22853.

[17] Elangovan, K., & Nalini, S. (2017). Plant disease classification using image segmentation and SVM techniques. *International Journal of Computational Intelligence Research*, 13(7), 1821-1828.

[18] Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., & Nandi, A. K. (2022). Medical image segmentation using deep learning: A survey. *IET image processing*, 16(5), 1243-1267.

[19] Hafiz, A. M., & Bhat, G. M. (2020). A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3), 171-189.

[20] Huang, Q., Luo, Y., & Zhang, Q. (2017). Breast ultrasound image segmentation: a survey. *International journal of computer assisted radiology and surgery*, 12, 493-507.

[21] Cardenas, C. E., Yang, J., Anderson, B. M., Court, L. E., & Brock, K. B. (2019, July). Advances in auto-segmentation. *In Seminars in radiation oncology* (Vol. 29, No. 3, pp. 185-197). WB Saunders.

[22] Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., & Sethi, A. (2017). A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7), 1550-1560.

[23] Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., ... & Rueckert, D. (2017). Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2), 384-395.

[24] Haque, I. R. I., & Neubert, J. (2020). Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, 18, 100297.

[25] Chowdhary, C. L., & Acharjya, D. P. (2020). Segmentation and feature extraction in medical imaging: a systematic review. *Procedia Computer Science*, 167, 26-36.

[26] Baumgartner, C. F., Koch, L. M., Pollefeys, M., & Konukoglu, E. (2018). An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. *In Statistical Atlases and Computational Models of the Heart*. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8 (pp. 111-119). Springer International Publishing.

[27] Khairuzzaman, A. K. M., & Chaudhury, S. (2017). Multilevel thresholding using grey wolf optimizer for image segmentation. *Expert Systems with Applications*, 86, 64-76.

[28] Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE access*, 9, 82031-82057.

[29] Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., & Asari, V. K. (2019). Recurrent residual U-Net for medical image segmentation. *Journal of medical imaging*, 6(1), 014006-014006.

[30] Ghosh, S., Das, N., Das, I., & Maulik, U. (2019). Understanding deep learning techniques for image segmentation. *ACM computing surveys* (CSUR), 52(4), 1-35.

[31] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7), 3523-3542.

[32] Roy, P., Goswami, S., Chakraborty, S., Azar, A. T., & Dey, N. (2017). Image segmentation using rough set theory: a review. *Medical Imaging: Concepts, Methodologies, Tools, and Applications*, 1414-1426.

[33] Seo, H., Badiei Khuzani, M., Vasudevan, V., Huang, C., Ren, H., Xiao, R., ... & Xing, L. (2020). Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state‐of‐art applications. *Medical physics*, 47(5), e148-e167.

[34] Hossain, M. D., & Chen, D. (2019). Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150, 115-134.

[35] Singh, V., & Misra, A. K. (2017). Detection of plant leaf diseases using image segmentation and soft computing techniques. *Information processing in Agriculture*, 4(1), 41-49.

[36] Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., & Rueckert, D. (2020). Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7, 25.

[37] Kuniki Imagawa & Kohei Shiomoto. (2024). Evaluation of effectiveness of pre-training method in chest X-ray imaging using vision transformer. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization(1),

[38] Wei Guo, Li Xu, Danyang Zhao, Dianqiang Zhou, Tian Wang & Xujing Tang.(2024).A Wind Power Combination Forecasting Method Based on GASF Image Representation and UniFormer. Journal of Marine Science and Engineering(7),1173-1173.

[39] Xianhui Peng, Chenchen Xu, Peng Zhang, Dandan Fu, Yan Chen & Zhigang Hu.(2024).Computer vision classification detection of chicken parts based on optimized Swin-Transformer. CyTA - Journal of Food(1),

[40] Zhonghui Guo, Dongdong Cai, Zhongyu Jin, Tongyu Xu & Fenghua Yu. (2025). Research on unmanned aerial vehicle (UAV) rice field weed sensing image segmentation method based on CNN-transformer. Computers and Electronics in Agriculture109719-109719.