Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 21(4): 5735–5761.
DOI: 10.3934/mbe.2024253
Received: 13 January 2024
Revised: 03 March 2024
Accepted: 05 March 2024
Published: 24 April 2024

Research article

Multi-phase features interaction transformer network for liver tumor segmentation and microvascular invasion assessment in contrast-enhanced CT

Wencong Zhang[1,4,†], Yuxi Tao[2,†], Zhanyao Huang[1], Yue Li[3], Yingjia Chen[1], Tengfei Song[2], Xiangyuan Ma[1,*] and Yaqin Zhang[2,*]

[1] Department of Biomedical Engineering, College of Engineering, Shantou University, Shantou, China
[2] Department of Radiology, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, China
[3] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[4] Department of Biomedical Engineering, College of Design and Engineering, National University of Singapore, Singapore

† These two authors contributed equally.

* **Correspondence:** Email: maxiangyuan@stu.edu.cn, zhyaqin@mail.sysu.edu.cn.

Abstract: Precise segmentation of liver tumors from computed tomography (CT) scans is a prerequisite step in various clinical applications. Multi-phase CT imaging enhances tumor characterization, thereby assisting radiologists in accurate identification. However, existing automatic liver tumor segmentation models did not fully exploit multi-phase information and lacked the capability to capture global information. In this study, we developed a pioneering multi-phase feature interaction Transformer network (MI-TransSeg) for accurate liver tumor segmentation and a subsequent microvascular invasion (MVI) assessment in contrast-enhanced CT images. In the proposed network, an efficient multi-phase features interaction module was introduced to enable bi-directional feature interaction among multiple phases, thus maximally exploiting the available multi-phase information. To enhance the model's capability to extract global information, a hierarchical transformer-based encoder and decoder architecture was designed. Importantly, we devised a multi-resolution scales feature aggregation strategy (MSFA) to optimize the parameters and performance of the proposed model. Subsequent to segmentation, the liver tumor masks generated by MI-TransSeg were applied to extract radiomic features for the clinical applications of the MVI assessment. With

Institutional Review Board (IRB) approval, a clinical multi-phase contrast-enhanced CT abdominal dataset was collected that included 164 patients with liver tumors. The experimental results demonstrated that the proposed MI-TransSeg was superior to various state-of-the-art methods. Additionally, we found that the tumor mask predicted by our method showed promising potential in the assessment of microvascular invasion. In conclusion, MI-TransSeg presents an innovative paradigm for the segmentation of complex liver tumors, thus underscoring the significance of multi-phase CT data exploitation. The proposed MI-TransSeg network has the potential to assist radiologists in diagnosing liver tumors and assessing microvascular invasion.

Keywords: liver tumor segmentation; multi-phase; contrast-enhanced CT; transformer

1. Introduction

Primary liver cancer, encompassing hepatocellular carcinoma and intrahepatic bile duct cancer, remains a critical global health challenge, ranked as a leading cause of cancer-related mortality. In 2020, more than 900,000 new cases of liver cancer were detected, and nearly 830,000 patients were killed by liver cancer [1]. The impact of liver cancer is still increasing, with an estimated over 1 million people to be affected by liver cancer in 2025 [2]. Early detection of liver cancer is crucial for successful treatments and can significantly improve the survival rate of patients [3,4]. Computed Tomography (CT) is one of the most widely used medical imaging modalities for liver cancer detection and diagnosis [5–7]. Studies have demonstrated that CT is a reliable method to detect early-stage liver cancer [8].

Accurate liver tumor segmentation in CT is a prerequisite step in various clinical applications, such as liver cancer diagnosis, microvascular invasion assessment [9,10], and preoperative planning for tumor resection and minimally invasive procedures [11,12]. However, conventional liver tumor segmentation approaches involve manual processes that are time-consuming, prone to human error, and susceptible to subjective personal experience. To tackle these issues, computer-aided diagnosis (CAD) technology has been developed to assist radiologists and oncologists to accurately interpret liver CT images. Initial approaches for the automatic delineation of tumors in medical imaging using CAD were anchored in rule-based algorithms, which utilized a set of predefined heuristics to distinguish between liver and tumor [13]. These methods encompass threshold-based methods [14–16], region-based methods [17], active contour-based methods [18], and clustering-based methods [19,20]. Despite their rapid execution and straightforward implementation, these techniques often fall short in accommodating the heterogeneity inherent to tumor presentations, thus consequently impeding their diagnostic efficacy.

In recent years, deep learning methods have demonstrated their superiority in solving various tasks in medical image analyses [21–24] and a variety of deep learning methods have been applied to the segmentation of the liver and liver tumors. Relatively early liver and liver tumors segmentation deep learning methods can be roughly classified into two categories.

In the first category, the methods used fully convolutional networks (FCN) as a backbone structure with various refinement strategies. For instance, Christ et al. [25] proposed a cascaded FCN to segment liver and liver tumors in CT images. This method fed segmented liver regions of interest (ROIs) into a second FCN to solely segment the liver tumors and then applied dense 3D conditional random fields (CRFs) to refine the segmentation results. Chlebus et al. [26] developed a 2D FCN with object-based post-processing to automatically segment liver tumors.

The second category encompasses the methods based on U-Net. U-Net was introduced by Ronneberger et al. [27], which was an updated FCN that extracted semantic or contextual information by contracting a path and used a symmetric expanding path to achieve accurate localization. Due to its straightforward structure and promising performance, various U-Net-like methods have emerged for liver and liver tumor segmentation. For instance, Li et al. [28] applied U-Net++ architecture with an attention-aware mechanism to segment the liver from CT images. Similarly, Huang et al. [29] applied the U-Net 3+ architecture for liver segmentation. Seo et al. [30] proposed a modified U-Net that incorporated a residual path into the skip connection of U-Net to improve the performance of liver tumor segmentation. Kushnure et al. [31] proposed a multi-scale liver tumor segmentation U-Net with a Res2Net module and a squeeze-and-excitation (SE) network to enhance the receptive field.

Despite the promising results achieved by these deep learning methods in automatic liver and liver tumor segmentation, most current liver tumor segmentation methods rely on single-phase CT images. This kind of approach may lead to unsatisfactory segmentation performance due to its inability to capture the complete morphology of the tumor, particularly under fuzzy tumor boundaries.

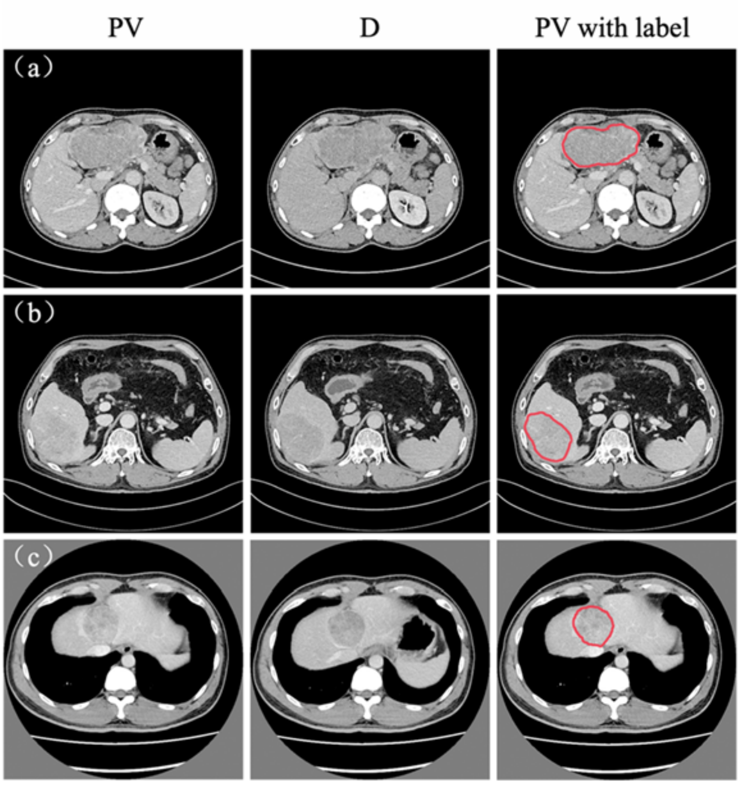

Figure 1. The portal venous phase (first column), delayed phase (second column) and their corresponding tumor labels (third column) from three patients. The red contour represents the radiologist’s hand outline of tumor boundary. PV and D represent portal venous phase and delayed phase.

In contrast, multi-phase images typically offer richer information about tumors [32], with distinct tumor regions differing from other tissues in either morphology or grayscale. A standard multi-phase

CT scanning protocol typically includes four phases: non-enhanced (NC), arterial (ART), portal vein (PV), and delayed (D) phases [33]. Generally, for liver CT scans, the maximum contrast between the liver tumor and the surrounding tissue appears in the PV phase (Figure 1(a)), which is the preferred phase for single-phase liver tumor segmentation [34]. However, the usage of the PV phase for tumor segmentation may not be appropriate in all cases. As shown in Figure 1(b),(c), the tumor boundaries appear fuzzy in the PV phase, while it has better contrast in the D phase. In clinical practice, radiologists simultaneously typically use multi-phase images to help them accurately identify the liver tumor boundary and diagnose liver cancer [35]. Therefore, liver tumor segmentation necessitates the use of multi-phase images to achieve precise and comprehensive results.

Several recent studies have attempted to improve the performance of automatic segmentation methods by combining the information from multi-phase images. Generally, multi-phase liver tumor segmentation methods utilize information from multi-phase images in the following three strategies: 1) the input-level fusion method (ILF) [36], which concatenated different phase images into a multi-dimensional map and used it as the input to the network; 2) the decision-level fusion method (DLF) [36,37], which independently processed each phase and then merged the output maps to obtain the final segmentation; and 3) the feature-level fusion method (FLF) [34,38,39], which combined features extracted from different phases and then decoded the combined multi-phases features to obtain the final segmentation result. While the first two strategies achieved better results than single-phase based methods, the third FLF strategy had the potential to make better use of multi-phase information; therefore, more research works exist. For example, Wu et al. [38] proposed a Modality Weighted U-Net (MW-UNet), which employed a phase-weighted sum rule to fuse features from multi-phase images at the decoder of the U-Net. Xu et al. [34] proposed a phase attention residual network (PA-ResSeg), which utilized a phase attention mechanism to exploit the features of the ART phase to improve the segmentation of PV phase.

While existing multi-phase based methods have demonstrated more accurate results than single-phase based methods, they predominantly relied on simplistic feature combination techniques such as concatenation and addition. This may hinder the full exploitation of cross-phase information, thus underscoring the necessity for an efficient cross-phase interaction mechanism. Moreover, the current paradigm of multi-phase based tumor segmentation predominantly revolves around Convolutional Neural Networks (CNNs), which mainly focus on extracting local information. However, the fixed receptive fields of CNNs constrain their ability to capture global contextual information crucial for distinguishing tumors from surrounding tissues.

Recognizing the inherent limitations of conventional approaches, we are motivated to explore the transformative potential of Transformer-based architectures in the realm of medical imaging. Originally devised for natural language processing tasks [40], the Transformer has garnered significant attention in medical imaging for its prowess in integrating global contextual information and modeling long-range dependencies. Several studies have underscored the superior performance of Transformer-based models in segmentation tasks compared to conventional approaches [41–43]. The self-attention mechanism lies at the heart of the Transformer, which is a fundamental feature that intricately associates different positions within a data sequence by harnessing query-key correlations. This unique characteristic proves particularly advantageous in managing cross-phase information and leveraging the global context of the segmentation.

The rising popularity of Transformer-based architectures can be attributed to their to more efficiently and effectively address the complex challenges of medical imaging [44–46]. He et al. [47]

combined Transformer and CNNs to enhance the capability of the model to extract global and local features. Hatamizadeh et al. [48] proposed Swin UNEt Transformers (Swin UNETR) for brain tumor segmentation, thus achieving top performance in related challenges. Zhu et al. [49] proposed a brain tumor segmentation method featuring a Swin Transformer-based semantic segmentation module, an edge detection module, and a feature fusion module, demonstrating great performance in brain tumor segmentation. Some methods such as X-Net [50] and Medical Transformer (MedT) [51] further emphasize the growing preference and effectiveness for Transformers in the field of medical imaging.

In particular, SegFormer introduces a hierarchically structured Transformer encoder and a lightweight MLP decoder, optimizing the architecture for semantic segmentation by unifying the generation of coarse and fine features without the need for positional encoding [52]. Consequently, employing a Transformer-based architecture not only offers significant promise in addressing the challenges associated with the efficient utilization of cross-phase information but also global contextual insights for liver tumor segmentation, thus addressing the sophisticated demands of the field.

In this study, we proposed a novel deep learning method, named the multi-phase feature interaction transformer segmentation network (MI-TransSeg), specifically designed for accurate liver tumor segmentation in contrast-enhanced CT images. The proposed method leveraged the D phase information to improve the segmentation performance in PV phase images. Comparative analyses with contemporary methodologies revealed that MI-TransSeg surpassed existing techniques in segmenting liver tumors. The salient contributions of this research are enumerated as follows:

- To fully utilize the available multi-phase information, we developed an efficient multi-phase features interaction module, thereby enabling bi-directional feature interaction among multiple phases.
- To effectively leverage global contextual information, we incorporated a hierarchical transformer-based encoder and decoder within the model, thus departing from the limitations of limited receptive fields in convolution blocks.
- We devised a multi-resolution scales feature aggregation strategy (MSFA) to optimize the parameters and performance of the proposed model.
- We applied the proposed segmentation model to the clinical application of the microvascular invasion (MVI) assessment. The results showed the promising potential of our method for clinical application.
- We performed a cross-center evaluation experiment to verify the generalizability of the proposed model.

2. Methods

The proposed multi-phase liver tumor segmentation method is comprised of three components, as illustrated in Figure 2. The first component is the Hierarchical Transformer Encoder, which takes two different phases of images as the input to generate high-resolution coarse features and low-resolution fine features for both phases. The second component is the Multi-phase Features Interaction, which enables the interaction of features among phases and produces Feature Tokens containing information from multiple phases. Finally, the Transformer Up-sampling Decoder receives the Feature Tokens from the Multi-phase Features Interaction module, along with an initially blank PV Feature Token. The decoder progressively up-samples the Feature Tokens while leveraging information from the Hierarchical Transformer Encoder and the tumor-related information from the learnable PV Feature

Token, ultimately predicting the final tumor segmentation results.

In the context of multi-phase liver tumor segmentation, the data and associated annotations can be categorized into two sets based on the image phase: $\mathcal{P} = \{X_P \in \mathbb{R}^{H\times W\times C}, Y_P \in \mathbb{R}^{H\times W}\}$, $\mathcal{D} = \{X_D \in \mathbb{R}^{H\times W\times C}\}$, where X_P and X_D represent the CT image in the PV phase and the D phase, respectively, with a resolution of $H \times W$ and C number of channels. As the liver tumor was annotated in the PV phase, the ground truth segmentation mask Y_P is available with a resolution of $H \times W$.

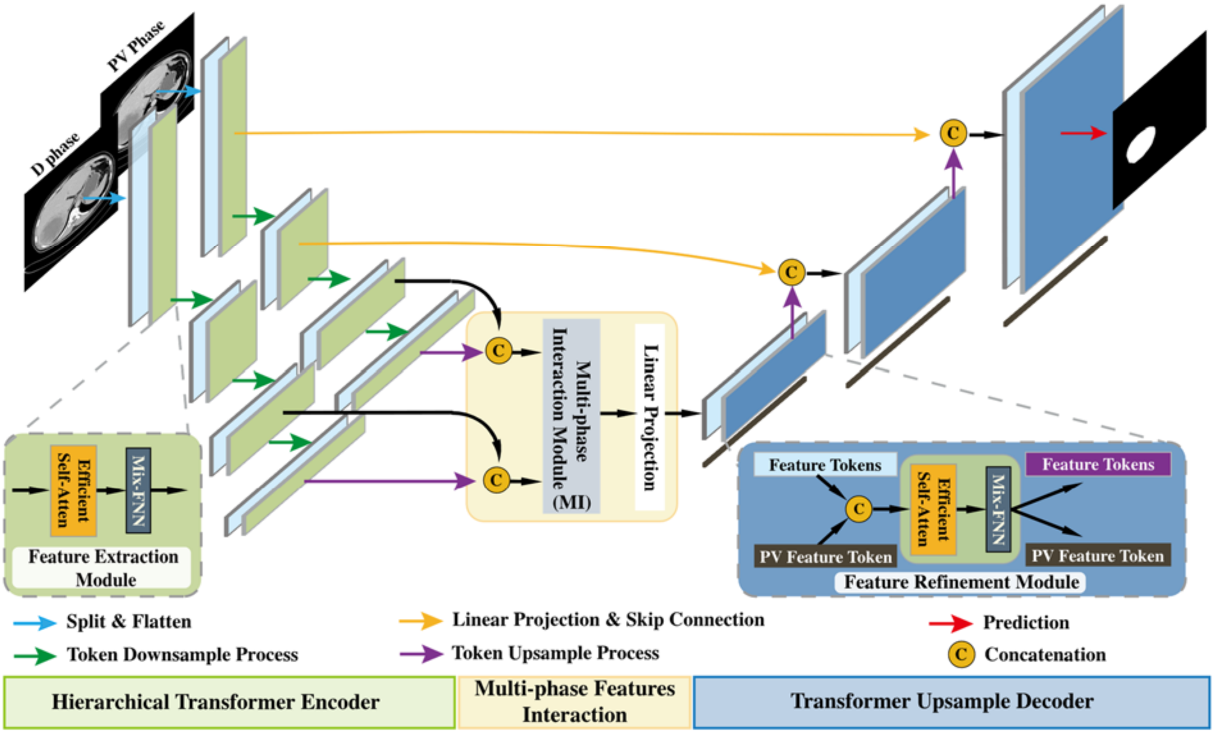
PV Phase
D phase
Efficient Self-Atten
Mix-FFN
Feature Extraction Module
Multi-phase Interaction Module (MI)
Linear Projection
C
Feature Tokens
Feature Tokens
PV Feature Token
PV Feature Token
Feature Refinement Module
Split & Flatten
Token Downsample Process
Linear Projection & Skip Connection
Token Upsample Process
Prediction
Concatenation
Hierarchical Transformer Encoder
Multi-phase Features Interaction
Transformer Upsample Decoder

Figure 2. Schematic diagram of the proposed MI-TransSeg method.

2.1. Hierarchical transformer encoder

We designed the Hierarchical Transformer Encoder, wherein we employed the pretrained SegFormer [52] model (mit-b4) as our backbone. The designed encoder aims to efficiently extract high-resolution coarse features and low-resolution fine features. Specifically, the Hierarchical Transformer Encoder is able to generate four resolution scale features that can bring more effective information for later multi-phase feature interaction and tumor prediction.

The designed encoder consists of three steps, following the flowchart depicted in Figure 2. The initial step in our encoder is the Split & Flatten process, which involves converting the input image into tokens. Next, the tokens are fed into the Feature Extraction Module to extract features. Note that the Mix-Feed Forward Network (Mix-FFN) is a component operation within the Feature Extraction Module. Finally, the extracted features are processed by the Token Down-sample Process to generate the next level of feature tokens. The details of the designed encoder are provided below and organized into two distinct modules: multi-resolution feature representation and the feature extraction module.

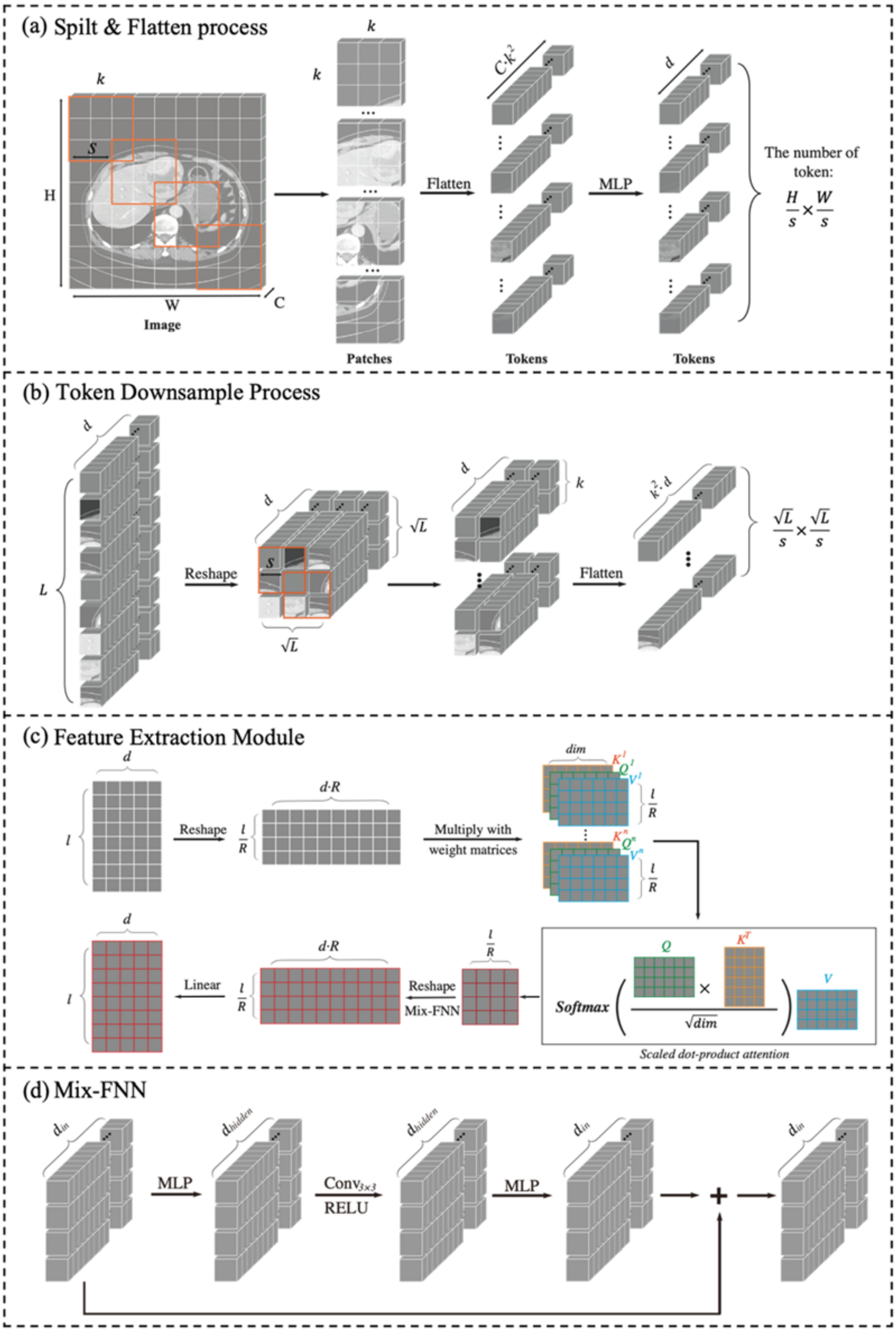

Figure 3. Four main processes in Hierarchical Transformer Encoder. (a) Illustration of Spilt & Flatten process, which converts the input images to tokens. (b) Illustration of Token Down-sample Process. L is the length of the tokens, k is the patch size, s is the stride, and p is zero padding sizes. (c) The Feature Extraction Module, which extracts the feature of the input tokens. (d) Mix-FNN in the Feature Extraction Module.

2.1.1. Multi-resolution feature representation

In contrast to the traditional Vision Transformer (ViT) [53] that was constrained to single-resolution feature maps, our enhanced module aims to produce multi-resolution features. Leveraging the hierarchical design, it extracts rich, CNN-like multi-resolution features from multiphase images, such as PV and D phase images, thus significantly improving the granularity and accuracy of liver tumor segmentation. In this Hierarchical Transformer Encoder, we generate four resolution features under the PV phase ($T_1^{PV}, T_2^{PV}, T_3^{PV}$, and T_4^{PV}) and four resolution features under the D phase ($T_1^{D}, T_2^{D}, T_3^{D}$, and T_4^{D}), respectively, as shown in Figure 2.

For the input PV phase X_P or D phase images X_D, we begin by converting these input images into tokens, which allows for the extraction of features using a self-attention mechanism. The process involves a Split & Flatten operation, illustrated in Figure 3(a). This operation divides the image into multiple patches with a resolution of $k \times k \times C$ and a stride of s. Zero-padding is also utilized to pad the image boundaries. Then, each patch is flattened into a token. Subsequently, all tokens are embedded together using a multi-layer perceptron (MLP) to generate the tokens T_1^{PV} and $T_1^{D} \in \mathbb{R}^{l_1 \times d}$, where $l_1 = \frac{H}{s} \times \frac{W}{s}$, d is embedding size.

To gradually down-sample the tokens and generate the remaining three resolution scales features of the PV and D phases (T_2^{PV}, T_3^{PV}, T_4^{PV}, T_2^{D}, T_3^{D} and T_4^{D}), the Token Down-sample Process (shown in Figure 3(b)) is designed and performed three times.

First, the Token Down-sample Process reshapes the tokens $T_i \in \mathbb{R}^{L \times d}$ into a square token map. Then, this reshaped token map is divided into smaller patches with a resolution of $k \times k \times d$ and a stride of s. Zero-padding is also utilized in this case. Subsequently, each smaller patch is flattened into a new token, generating new token features $T_{i+1} \in \mathbb{R}^{\frac{L}{s^2} \times kd}$.

In this study, for the four resolution scales, the patch sizes are set to k = [7, 3, 3, 3], the strides are set to s = [4, 2, 2, 2], and the zero padding sizes are set to p = [3, 1, 1, 1]. Through these operations, we obtain four resolution scales features of the PV and D phases, donated as $T_1^{phase} \in \mathbb{R}^{l_1 \times d}$, $T_2^{phase} \in \mathbb{R}^{l_2 \times d}$, $T_3^{phase} \in \mathbb{R}^{l_3 \times d}$, and $T_4^{phase} \in \mathbb{R}^{l_4 \times d}$, where $phase = PV\ or\ D, l_1 = \frac{H}{4} \times \frac{W}{4}$, $l_2 = \frac{H}{8} \times \frac{W}{8}$, $l_3 = \frac{H}{16} \times \frac{W}{16}$, $l_4 = \frac{H}{32} \times \frac{W}{32}$ and d is the embedding size.

2.1.2. Feature extraction module

To extract the tumor-related global contextual information at each resolution scale, we implement the self-attention mechanism following the Spilt & Flatten process and the Token Down-sample Process, as shown in Figure 2. Since the convention multi heads self-attention is extremely computational and complex when facing tokens with large resolutions, we adopt a sub-module called efficient self-attention in this study to reduce the computational complexity.

Figure 3(c) illustrates the Feature Extraction Module with an efficient self-attention mechanism, which incorporates a sequence reduction process initially introduced in [52,53] to reduce the number of the tokens so as to reduce the computational complexity. First, we reshape the $T_i \in \mathbb{R}^{l_i \times d}$ to $T_i' \in \mathbb{R}^{\frac{l_i}{R} \times d \cdot R}$, where R is the reduction ratio. Then, the tokens T_i' is multiplied with the weighted matrices to generate three kinds of matrices; the *K* (key), *Q* (query), and *V* (value). These matrices are used to calculate the scaled-dot product attention score [40]. Following this, we reshape the scaled-dot product attention score and feed it into the Mix-Feed Forward Network (Mix-FFN), which accounts for zero padding's impact on the leak location information and incorporates a 3 × 3 Conv [52]. Placed after the self-attention, the Mix-FFN further locally processes information for each token. The flowchart of the Mix-FFN is illustrated in Figure 3(d). Finally, we use a linear layer to reshape the tokens $T_i'' \in \mathbb{R}^{\frac{l_i}{R}\times d\cdot R}$ back to tokens $T_i''' \in \mathbb{R}^{l_i\times d}$.

For given tokens $T \in \mathbb{R}^{l_i\times d}$, the computational complexity can be represented as $O(l_i^2)$. By introducing the sequence reduction process with the reduction ratio R, we are able to reduce the computational complexity of self-attention from $O(l_i^2)$ to $O(\frac{l_i^2}{R^2})$. In our method, we set the reduction ratio R as [8, 4, 2, 1] from the first resolution scale to the fourth resolution scale.

2.2. Multi-phase features interaction

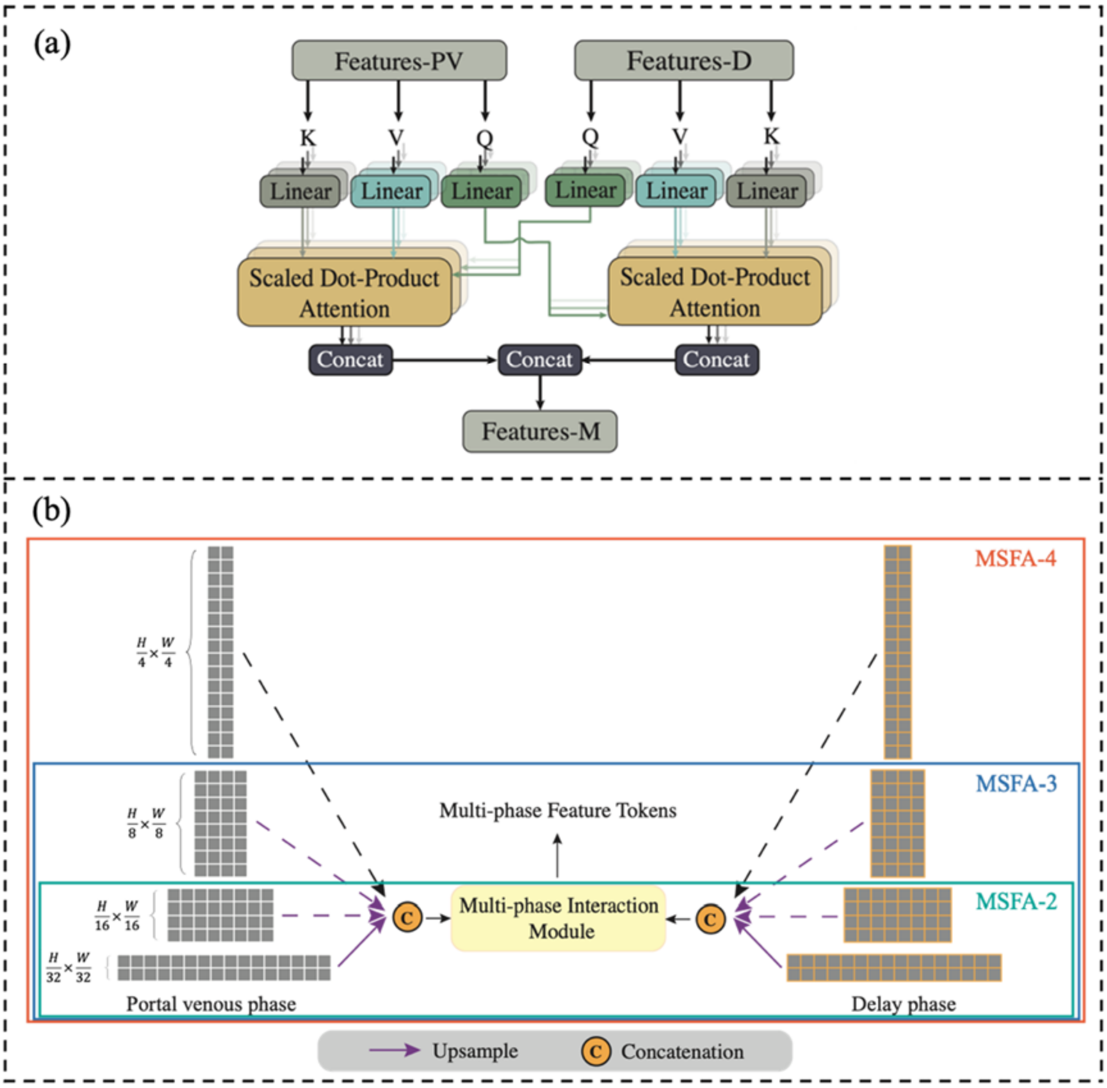
(a)
Features-PV
Features-D
K
V
Q
Q
V
K
Linear
Linear
Linear
Linear
Linear
Linear
Scaled Dot-Product Attention
Scaled Dot-Product Attention
Concat
Concat
Concat
Features-M
(b)
MSFA-4
MSFA-3
MSFA-2
$\frac{H}{4}\times\frac{W}{4}$
$\frac{H}{8}\times\frac{W}{8}$
$\frac{H}{16}\times\frac{W}{16}$
$\frac{H}{32}\times\frac{W}{32}$
Multi-phase Feature Tokens
C
Multi-phase Interaction Module
C
Portal venous phase
Delay phase
Upsample
C Concatenation

Figure 4. (a) Illustration of our proposed multi-phase interaction module for liver tumor segmentation. Features-PV, Features-D, and Features-M are the features of PV phase, D phase, and multi-phase features after cross-phase interaction. (b) Illustration of our proposed multi-resolution scales feature aggregation (MSFA) strategy, where MSFA-2, MSFA-3, MSFA-4 represent utilizing the features of two, three, and four resolution scales respectively.

Motivated by the feature-level fusion methods that facilitate bi-directional information exchange [22,54], we modify the self-attention mechanism [40] to achieve a bi-directional cross-phase information interaction, thus proposing a Multi-phase Features Interaction Module, as shown in Figure 4(a). Importantly, leveraging the efficiency of our Hierarchical Transformer Encoder to extract the liver tumor features across four resolution scales, we aim to fully harness the tumor features from the PV and D phases at multiple resolutions. Thus, we introduce a multi-resolution scales feature aggregation (MSFA) strategy, depicted in Figure 4(b), which aggregates features at two or more resolution levels. Different from the traditional single-resolution strategies [34,38,39] that only fuse the features at the lowest resolution level, the proposed strategy aims to improve the segmentation performance of liver tumors by aggregating features at multiple resolution scales.

2.2.1. Multi-phase features interaction module

Our proposed Multi-phase Features Interaction Module is modified from a self-attention mechanism, and it implements cross-phase communication by cross-stream manner. As illustrated in Figure 4(a), the proposed module achieves multi-phase bi-directional communication by computing the relationship between two phases. The calculated cross-phase relationships are subsequently aggregated to generate multi-phase feature tokens. The details of the Multi-phase Features Interaction Module are provided as follows.

Given the input features F^{PV} and F^{D}, we embed them into queries $Q_{PV} \in \mathbb{R}^{l \times d}$, $Q_{D} \in \mathbb{R}^{l \times d}$, keys $K_{PV} \in \mathbb{R}^{l \times d}$, $K_{D} \in \mathbb{R}^{l \times d}$, and values $V_{PV} \in \mathbb{R}^{l \times d}$, $V_{D} \in \mathbb{R}^{l \times d}$, respectively, by multiplying the weight matrices. Then, the queries, keys, and values of the two phases are projected h times with different, learned linear projections to $d_q, d_k,$ and d_v dimensions, respectively [40]. The Q_{PV}, K_D, and V_D are inputted into the Scaled Dot-Production Attention to calculate the long-range dependence values between the PV and D features. Accordingly, Q_D, K_{PV}, and V_{PV} are also inputted to another Scaled Dot-Production Attention to calculate the long-range dependence value in the reverse direction. Finally, these calculated values are concatenated and projected to yield multi-phase feature tokens.

2.2.2. Multi-resolution scales feature aggregation strategy

The proposed Multi-resolution Scales Feature Aggregation Strategy (MSFA) aims to enhance the performance of segmentation by aggregating the tumor features in multi-resolution levels extracted by the Hierarchical Transformer Encoder. As shown in Figure 4(b), given the features from two phases at four resolution scales, aggregating features from two resolution scales (MSFA-2) involves up-sampling and concatenating features from the fourth resolution level ($\frac{H}{32} \times \frac{W}{32}$) to the third resolution level ($\frac{H}{16} \times \frac{W}{16}$) for each phase. Then, the aggregated features from each phase are input to the Multi-phase Features Interaction Module to obtain the multi-phase feature tokens. The Up-sample process in MSFA strategy can be viewed as a linear projection. First, it reshapes tokens into an image first, then applies a linear projection to up-sample the image to the desired resolution. Last, we split the image into tokens again. After the Up-sample process, all tokens are projected to the same embedding size, which makes sure they can be concatenated.

To determine the optimal model for liver tumor segmentation, we compare the models using MSFA with a different number of resolution scales. When using features at three resolution scales (MSFA-3), we need to up-sample the features at the fourth and third resolution scales to the second resolution ($\frac{H}{8} \times \frac{W}{8}$), and then input them into the Multi-phase Features Interaction Module to obtain the multi-phase feature tokens. In the case of using four resolution levels (MSFA-4), we need to up-sample the features at the fourth, third, and second resolution scales to the first resolution scale ($\frac{H}{4} \times \frac{W}{4}$) and input them into the Multi-phase Features Interaction Module to obtain the multi-phase features.

In summary, to implement each Multi-resolution Scales Feature Aggregation Strategy, we follow a structured process. First, we up-sample the features of all phases associated with the strategy to the highest resolution. Next, we concatenate these features from resolution levels within each phase. Finally, the concatenated features are inputted into the Multi-phase Features Interaction Module. The highest resolution levels corresponding to MSFA-2, MSFA-3, and MSFA-4 are ($\frac{H}{16} \times \frac{W}{16}$), ($\frac{H}{8} \times \frac{W}{8}$), and ($\frac{H}{4} \times \frac{W}{4}$), respectively.

2.3. Transformer up-sample decoder

For the liver tumor segmentation task, achieving finer tumor boundaries is crucial for accurate tumor diagnoses and subsequent applications. However, directly up-sampling low-resolution tokens to predict segmentation results often fails to produce high-quality outcomes due to the insufficient information contained within these tokens [43]. To address this challenge, we propose a Transformer Up-sample Decoder equipped with a hierarchical feature recovery module and introduce a PV feature token to refine the liver tumor predictions. As shown in Figure 2, the Transformer Up-sample Decoder is mainly composed of Hierarchical Feature Recovery (Token Up-sample Process and skip connection) and the Feature Refinement Module, and the following will introduce how the Transformer Up-sample Decoder achieves our goal.

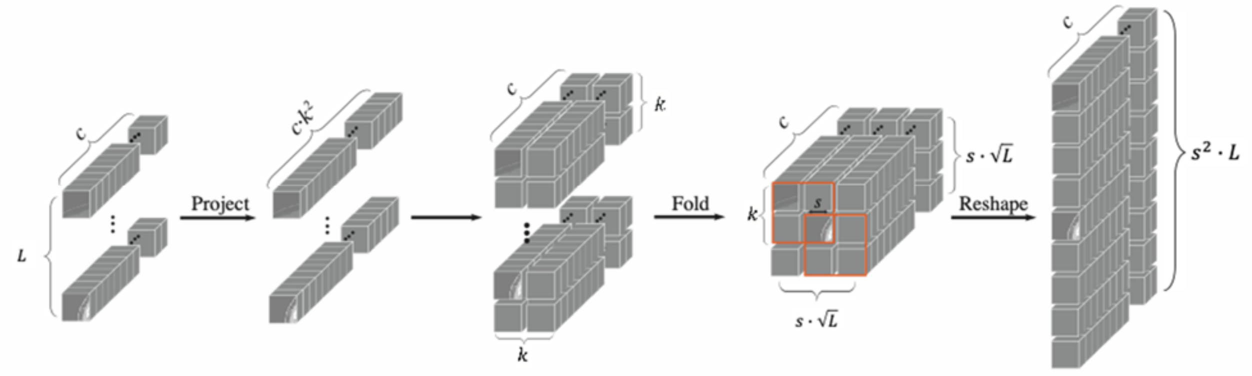

Figure 5. Illustration of Token Upsample Process. L is the length of the tokens, k is the patch size, s is the overlapping strike, and p is zero padding sizes.

2.3.1. Hierarchical feature recovery

The hierarchical feature recovery design gradually up-samples tokens to enhance the tumor feature recovery, consisting of a Token Up-sample Process and a Skip connection, as illustrated in Figure 2. In the MSFA-2 strategy, given multi-phase feature tokens T_3^M output from Multi-phase Features Interaction Module, we use the Token Up-sample Process to up-sample the tokens T_3^M to tokens T_2^M, corresponding to the second resolution scale. Subsequently, the tokens T_2^M are concatenated with the low-level tokens T_2^{PV} from the Hierarchical Transformer Encoder. We repeat the Token Up-sample Process and skip connection to progressively up-sample the feature tokens to the first resolution scale ($\frac{H}{4} \times \frac{W}{4}$). Finally, the up-sampled feature tokens are input to the Prediction process for predicting the final liver tumor mask. The Prediction process also contains a Token Up-sample Process for up-sampling tokens to the full resolution level ($H \times W$) and then the tokens are reshaped into an image to obtain the final liver tumor mask.

The Token Up-sample Process, which can be considered as an inverse Token Down-sample Process, is depicted in Figure 5. Given a sequence of multi-phase feature tokens $T_i^M \in \mathbb{R}^{l_i \times c}$, we first project the input tokens T_i^M to expand their embedding dimension from c to ck^2. Next, each token is seen as a $k \times k$ patch. Then, we fold these patches into an image, keeping the s stride and p zero-padding with the neighboring patches. The dimension of the image is $s\sqrt{L} \times s\sqrt{L} \times c$, where $c = 64$. Finally, the image is reshaped to new tokens $T_{i-1} \in \mathbb{R}^{s^2 L \times c}$. Among the three Token Up-sample Processes, the patch size is set to $k =$ [3,3,7], $s =$ [2,2,4], and $p =$ [1,1,3].

2.3.2. Feature refinement module with PV feature token

The Feature Refinement Module is modified from the Feature Extraction Module defined in Section 2.1.2. In this module, a learnable task-related token, called PV Feature Token, is incorporated to refine the Feature Tokens produced by the Multi-phase Features Interaction module. The design of the PV Feature Token is inspired by the existing Transformer methods, which utilize either a learnable token [53,55] or a task-related token [56] to improve the prediction accuracy.

The Feature Refinement Module enables the interaction between the PV Feature Token and the Feature Tokens. During the interaction between the PV Feature Token and the Feature Tokens, the PV Feature Token can learn tumor-related embedding, which can be used to further refine the Feature Tokens during the interaction in the next Feature Refinement Module. The feature refinement module is placed before each Token Up-sample Process, as shown in Figure 2.

In detail, within the Feature Refinement Module, the Feature Tokens $T_i^M \in \mathbb{R}^{l_i \times d}$ are initially combined with the PV feature token $T_{PV} \in \mathbb{R}^{1 \times d}$. These concatenated tokens $T^c \in \mathbb{R}^{L_i \times d}, L_i = l_i + 1$ then undergo processing through Efficient Self-attention and Mix-FNN mechanisms. Subsequently, the processed tokens are separated back into Feature Tokens and a PV Feature Token. The Feature Tokens undergo further processing via the Token Up-sample Process, while the PV Feature Token inputs the Feature Refinement Module in the next resolution into scale.

3. Experiments and results

3.1. Dataset and preprocessing

The clinical multi-phase contrast-enhanced CT abdominal dataset, which included 164 patients with liver tumor, was collected from Sun Yat-Sen University's Third Affiliated Hospital, Guangzhou, China (108 patients), and Sun Yat-Sen University's Fifth Affiliated Hospital, Zhuhai, China (56 patients). Each patient in the dataset contained multi-phase CT images, including non-enhanced phase (NC), arterial phase (ART), portal vein phase (PV) and delayed phase (D). Each phase has 34 to 679 axial slices with thickness from 2 to 5 mm and pixel spacing from 0.63 to 0.87 mm. This multi-center dataset has obtained ethics approval and consent from the Medical Ethics Committee of the Fifth Affiliated Hospital of Sun Yat-sen University for retrospective usage (Approval No: No. [2023] K20-1). All participants' consent were waived by the Medical Ethics Committee of the Fifth Affiliated Hospital of Sun Yat-sen University and Third Affiliated Hospital of Sun Yat-sen University.

An experienced radiologist drew the liver tumor masks for all the patients in the PV phase by simultaneously observing multi-phase CT images. In this study, the hand segmented liver tumor masks were verified by two other experienced radiologists and then used as a reference standard. In addition, in order to evaluate the performance of the MVI assessment, histopathology examination results for all the patients were collected, which contain information about the MVI category of the patients.

All models were validated under a cross validation scheme. Concretely, we randomly split the dataset into 70% (114 patients) and 30% (50 patients) for training and testing, respectively. In the training process, 80% patients were randomly selected to train the model, while the remaining 20% patients were used to validate the model.

For data pre-processing, we truncated the image intensity values of all scans of [-80, 220] HU and performed the normalization on these scans to enhance the contrast in the liver and tumor region. To prevent potential overfitting problems and improve the robustness of the model, the data was augmented by commonly used augmentation methods, including randomly shifting, rotating, and scaling.

3.2. Experiment setup

The proposed MI-TransSeg was achieved based on Python 3.8 and PyTorch 1.8.2 [57]. We trained our model on an NVIDIA GeForce RTX 2080 Ti with 11-GB VRAM. On the model training, the initial learning rate was set as 0.0001, and the Adam [58] gradient descend with momentum was used to optimize the model. The decaying learning rate strategy was applied in this study, which reduces the learning rate by a factor of 10 when the training loss value stops falling for 10 epochs. The total loss is the sum of the cross-entropy loss of the masks at the four resolution levels, and the cross-entropy loss at each resolution level is multiplied by a weight (λ_{ce}). The weights are 1, 0.8, 0.8, 0.5 from the first resolution level to the fourth, respectively. The overall training configuration is shown in Table 1.

Table 1. Training configuration details.

Parameter	Value
Learning rate	1×10^{-4}
Decay factor	10
Optimizer	Adam
Epochs	150
Loss function	$\mathcal{L}_{loss} = \lambda_{ce1}\mathcal{L}_{ce1} + \lambda_{ce2}\mathcal{L}_{ce2} + \lambda_{ce3}\mathcal{L}_{ce3} + \lambda_{ce4}\mathcal{L}_{ce4}$

3.3. Evaluation metric

We evaluated the segmentation performance of the network by comparing the ground truth with the segmentation result using evaluation metrics including the Dice similarity coefficient (DSC), sensitivity, the absolute relative volume difference (ARVD), and the average symmetric surface distance (ASSD). The DSC, sensitivity, and ARVD are volumetric size similarity metrics, while ASSD is the surface distance metric.

1) Dice similarity coefficient (DSC)

The DSC is used to measure the overlap between the prediction and the ground truth, and is expressed as follows Eq (1):

$$DSC(G,P) = \frac{2|G \cap P|}{|G| + |P|} \times 100\%, \tag{1}$$

where G denotes the ground truth and P represents the prediction result.

2) Sensitivity

The sensitivity, also known as the True Positive Rate, reflects the proportion of the correctly predicted pixels to the true tumor pixels. The Sensitivity is calculated by Eq (2).

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100\%, \tag{2}$$

where TN is a true positive and FN is a false negative.

3) The absolute relative volume difference (ARVD)

The ARVD represents the relative difference between the ground truth and the predicted volumes of a segmented structure, which is normalized by the ground truth volume, as shown in Eq (3).

$$ARVD = \left| \frac{|P| - |G|}{|G|} \right| \times 100\%, \tag{3}$$

where G is the ground truth and P is the predicted results.

4) The average symmetric surface distance (ASSD)

The ASSD measures the average distance between the surfaces of the ground truth and the predicted result. Equation (4) calculates the shortest distance of a random voxel v to the surface voxels of the ground truth S_G, which is calculated by Eq (5).

$$d(v, S_G) = \min_{v_G \in S_G} \parallel v - v_G \parallel, \tag{4}$$

$$ASSD = \frac{1}{N(G) + N(P)} \left(\sum_{v \in S_P} d(v, S_G) + \sum_{v \in S_G} d(v, S_P) \right), \tag{5}$$

where S_P represents the surface voxels of the predicted result, and $N(G)$, $N(P)$ represent the number of voxels of the ground truth and the predicted result, respectively.

The DSC, Sensitivity, and ARVD are expressed as percentages, while the ASSD is measured in millimeters (mm). A segmentation score of 100% is considered the best for the DSC and sensitivity, while 0% is the best score for the ARVD. For the ASSD, the best score is 0 mm.

3.4. Comparison with the state-of-the-arts

To validate the superiority of our proposed MI-TransSeg, five state-of-the-art models are considered as baseline methods in our comparison study. The baseline methods can be classified into two categories:

1) Typical single-phase medical image segmentation methods (U-Net, TransUnet and SegFormer)
2) Feature-level fusion multi-phase segmentation methods (MW-Unet, PA-ResSeg)

The results for different methods are summarized in Table 2 and Figure 6. It is obvious that multi-phase-based methods (MW-Unet, PA-ResSeg, and our proposed method) outperformed single-phase methods (U-Net, TransUnet and SegFormer), demonstrating that the FLF segmentation methods utilizing multi-phase information were able to produce more precise results. Meanwhile, our proposed MI-TransSeg achieved the best performance among the competing models, having the highest DSC ($P < 0.05$) and Sensitivity and the lowest ARVD scores ($P < 0.05$) and ASSD. It improves the DSC, Sensitivity, ARVD, and ASSD to 91.36%, 94.06%, 5.96%, and 2.95%, respectively. Furthermore, the shapes and sizes of our method-predicted liver tumors were the most accurate compared to the ground truth, as shown in Figure 6(a).

For further verification of whether our Transformer based method has an improved global information extraction capability, we visualized the effective acceptance field (ERF) of the six methods according to the method mentioned in this article [59], the results of which are shown in Figure 6(b). We observed that the ERFs of U-Net, TransUNet, MW-Unet, and PA-ResSeg are relatively concentrated. SegFormer generated local attention while was also capable of capturing certain global information. MI-TransSeg demonstrated stronger local attention as well as a wider global attention since the ERF of MI-TransSeg had the widest range and the strongest intensity. Thus, our novel proposed MI-TransSeg method delivered the best performance.

Table 2. The results of different methods ($MeanValue \pm StandardDeviation$) on the multi-phase contrast-enhanced liver tumor CT dataset.

Methods	DSC (%)	Sensitivity (%)	ARVD (%)	ASSD (mm)
Single-phase				
U-Net	86.63 ± 1.64	93.23 ± 1.36	7.41 ± 1.55	5.06 ± 1.90
TransUnet	87.23 ± 1.96	92.16 ± 1.42	7.01 ± 1.53	5.53 ± 3.10
SegFormer	89.23 ± 1.59	92.54 ± 1.43	7.90 ± 1.58	3.87 ± 1.37
Multi-phase				
MW-UNet	89.58 ± 1.83	91.55 ± 1.73	8.92 ± 1.81	4.52 ± 2.35
PA-ResSeg	89.78 ± 1.67	93.58 ± 1.45	6.85 ± 1.54	4.05 ± 1.60
Ours	91.36 ± 1.30	94.06 ± 1.23	5.96 ± 1.14	2.95 ± 1.01

Note: *T-test was used to examine the difference in performance between the state-of-the-art methods and our method(Ours vs U-Net, Ours vs TransUnet, Ours vs SegFormer, Ours vs MW-UNet, Ours vs PA-ResSeg). P-values of DSC are all P < 0.05; P-values of Sensitivity are P = 0.07, P < 0.05, P < 0.05, P < 0.05, and P = 0.28; P-values of ARVD are all P < 0.05; P-values of ASSD are P < 0.05, P < 0.05, P = 0.11, P < 0.05, and P < 0.05.

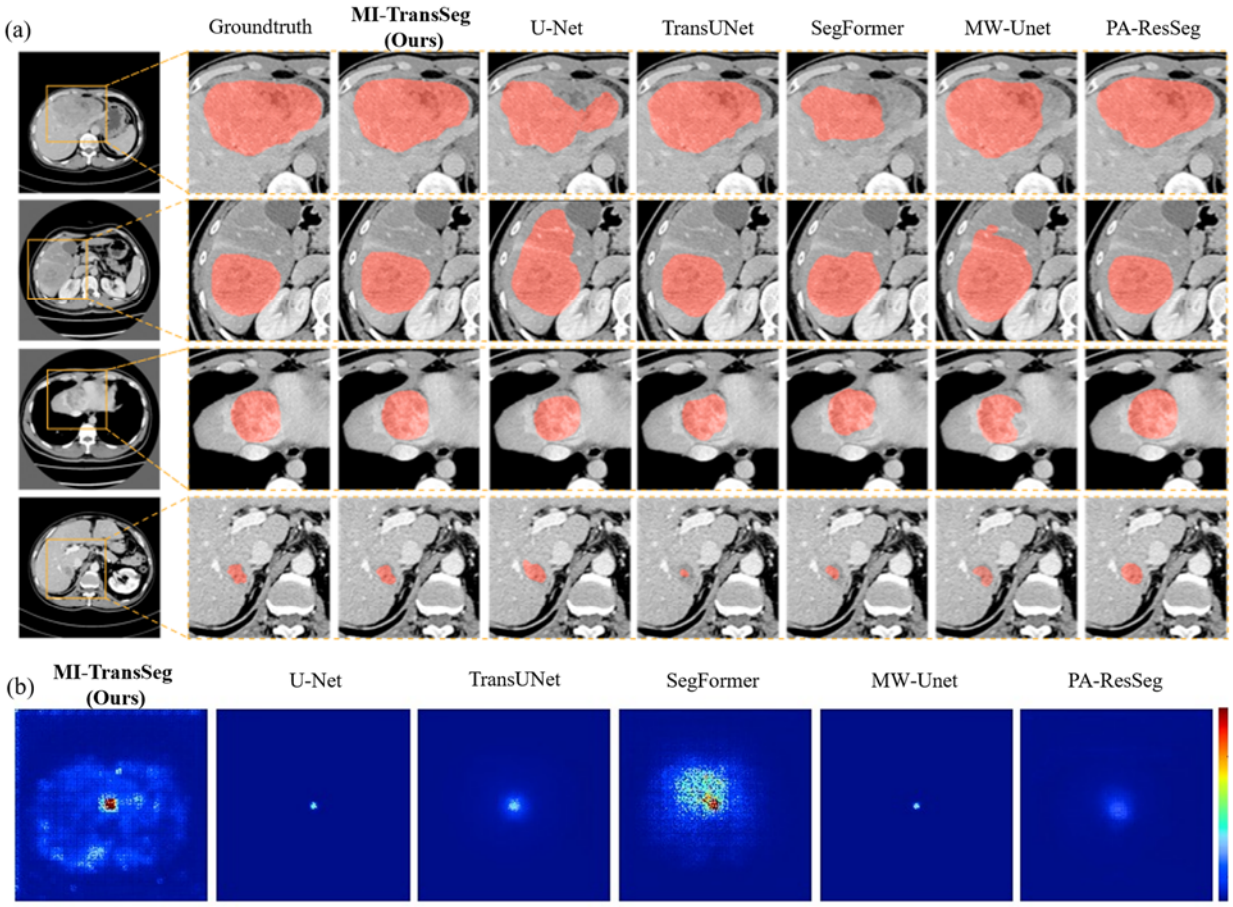

Figure 6. Visual examples of the performance of different methods in four different sizes of tumors**.** (a) segmentation results of different methods. The details of segmentation results are zoomed for better comparison. (b) the Effective Receptive Field (ERF) of different methods on the liver tumor dataset (calculated at 100 images).

3.5. Ablation study

We conducted an ablation study to analyze the influence of different designs and components in our proposed MI-TransSeg.

3.5.1. The effectiveness of multi-resolution scales feature aggregation strategy (MSFA)

As previously mentioned, we proposed a multi-phase feature interaction approach to effectively aggregate complementary information from multi-phase images, thereby enhancing finer segmentation details. The purpose of this ablation experiment is to assess the impact of different complementary information utilization strategies on the segmentation results using different numbers of resolution scales. Four models have been compared: MSFA-1, MSFA-2, MSFA-3, and MSFA-4.

1) MSFA-1: Using only the features of a single-resolution scale $(\frac{H}{32} \times \frac{W}{32})$;

2) MSFA-2: Aggregating the features of two resolution scales $(\frac{H}{32} \times \frac{W}{32})$ $(\frac{H}{16} \times \frac{W}{16})$;

3) MSFA-3: Aggregating the features of three resolution scales $(\frac{H}{32} \times \frac{W}{32})$ $(\frac{H}{16} \times \frac{W}{16})$ $(\frac{H}{8} \times \frac{W}{8})$;

4) MSFA-4: Aggregating the features of four resolution scales $(\frac{H}{32} \times \frac{W}{32})$ $(\frac{H}{16} \times \frac{W}{16})$ $(\frac{H}{8} \times \frac{W}{8})$ $(\frac{H}{4} \times \frac{W}{4})$.

Table 3 shows the results of the models using different numbers of resolution scales. In our liver tumor task segmentation, the MSFA strategies with the more desirable results aggregate features at two resolution scales (MSFA-2) and aggregate features at three resolution scales (MSFA-3), which improve the segmentation performance and increase the DSC from 90.07% to 91.36% and 91.40%. MSFA-2 and MSFA-3 achieved similar performances, though aggregating a greater number of scales resulted in a higher computational cost; therefore, MSFA-2 was used in this study. These results demonstrated the superiority of our proposed multi-resolution scales feature merging strategy.

Table 3. Performance of the models with different multi-resolution scales feature aggregation strategy (MSFA). The segmentation performance ($MeanValue \pm StandardDeviation$) is tested on the multiphase contrast-enhanced liver tumor CT dataset.

Methods	DSC (%)	Sensitivity (%)	ARVD (%)	ASSD (mm)
MSFA-1	90.07 $\pm$ 1.30	94.00 $\pm$ 1.25	6.35 $\pm$ 1.37	4.55 $\pm$ 2.06
MSFA-2 (MI-TransSeg)	91.36 $\pm$ 1.30	94.06 $\pm$ 1.23	5.96 $\pm$ 1.14	2.95 $\pm$ 1.01
MSFA-3	91.40 $\pm$ 1.40	93.66 $\pm$ 1.36	5.89 $\pm$ 1.21	3.15 $\pm$ 0.94
MSFA-4	89.90 $\pm$ 1.55	93.16 $\pm$ 1.25	6.30 $\pm$ 1.24	3.76 $\pm$ 1.42

3.5.2. The effectiveness of each important component in the proposed MI-TransSeg

To validate the efficacy of each crucial component in our proposed method, namely the transformer up-sample decoder (TUD), PV feature token (PVFT), and Multi-phase features interaction module (MI), we conducted ablation studies on our clinical multi-phase contrast-enhanced CT liver tumor dataset. The ablation study consisted of four experiments:

1) Baseline: We removed the TUD, PVFT, and MI from our MI-TransSeg to establish a baseline. Specifically, we extracted the PV phase feature tokens and the D phase feature tokens at 1/32, 1/16, 1/8, and 1/4 resolution scales using the hierarchical transformer encoder. Then, the features were directly concatenated to predict the tumor segmentation results using MLP.
2) "+TUD": We deployed the TUD after the hierarchical transformer encoder to validate the effectiveness of our proposed transformer up-sample decoder.
3) "+TUD + PVFT": We further applied PVFT to the TUD to verify the effectiveness of the PV feature token.
4) "+TUD + PVFT + MI": To verify the effectiveness of the multi-phase feature interaction module, we substituted MI for the concatenation method after the hierarchical transformer encoder.

The results of these four experiments are presented in Table 4, demonstrating that all three components have a positive contribution to the performance improvement of MI-TransSeg. Compared to directly using low-resolution features to predict tumors (Baseline), our proposed transformer up-sample decoder demonstrated an improved performance, with the DSC increasing from 83.23% to 86.24%. Additionally, these results indicated that leveraging both low-resolution and high-resolution features to gradually recover full-resolution features could result in finer tumor segmentation results, thus improving the ASSD from 9.21 to 5.50 mm. With the inclusion of the PV feature token, the DSC and sensitivity increased from 86.24 and 89.22% to 90.19% and 93.77%, and the ARVD and ASSD decreased from 9.71% and 5.50 mm to 7.16% and 3.54 mm, respectively, demonstrating that the PV feature token can enhance the segmentation performance by interacting with the main feature token to learn the image-related task embedding. Moreover, the performance of the model further improved after applying the multi-phase feature interaction module, with the DSC and sensitivity both increasing to 91.36% and 94.06%, respectively, and the ARVD and ASSD both decreasing to 5.96% and 2.95 mm. These results suggested that the multi-phase feature interaction module can more effectively utilize multi-phase complementary information compared to directly concatenating multi-phase features.

Table 4. The effectiveness of each component in our proposed method ($\mathrm{MeanValue} \pm \mathrm{StandardDeviation}$). "TUD" denotes transformer upsample decoder. "PVFT" represents PV feature token, and "MI" stands for multi-phase feature interaction module.

Methods	DSC (%)	Sensitivity (%)	ARVD (%)	ASSD (mm)
Baseline	83.23 ± 2.10	90.41 ± 1.84	8.72 ± 2.25	9.21 ± 2.00
+TUD	86.24 ± 1.93	89.22 ± 2.02	9.74 ± 1.97	5.50 ± 2.16
+TUD + PVFT	90.19 ± 1.57	93.77 ± 1.31	7.16 ± 1.46	3.54 ± 0.96
+TUD + PVFT+ MI	91.36 ± 1.30	94.06 ± 1.23	5.96 ± 1.14	2.95 ± 1.01

3.5.3. The effectiveness of multi-phase strategy

To examine the impact of multi-phases on the model performance, we conducted the following experiments: 1) TransSeg (PV), where we removed the MI from our MI-TransSeg and used the PV phase as input; 2) MI-TransSeg (PV&D), where we employed complementary information from the D phase to aid liver tumor segmentation in the PV phase images; and 3) MI-TransSeg (PV&D&ART), where we employed complementary information from the D phase and ART phase to aid liver tumor segmentation in PV phase images.

Table 5. Performance comparison between models using single-phase (PV) and multi-phase (PV&D and PV&D&ART) $(MeanValue \pm StandardDeviation)$. PV, D, and ART represent arterial phase, delayed phase, and arterial phase.

Methods	DSC (%)	Sensitivity (%)	ARVD (%)	ASSD (mm)
TransSeg (PV)	89.51 ± 1.82	94.05 ± 1.32	6.28 ± 1.47	4.38 ± 1.75
MI-TransSeg (PV&D)	91.36 ± 1.30	94.06 ± 1.23	5.96 ± 1.14	2.95 ± 1.01
MI-TransSeg (PV&D&ART)	90.41 ± 1.55	94.78 ± 1.15	5.33 ± 1.33	3.75 ± 1.47

Table 5 presents the segmentation results derived from single-phase (PV) and multi-phase (PV&D and PV&D&ART) images. The results suggest that the use of multi-phase images substantially enhances the results of tumor segmentation, surpassing those achieved through the use of single-phase images. In addition, we find that using three-phase images does not effectively improve the performance of segmentation, but instead reduces the DSC from 91.36% to 90.41%. Therefore, we believe that two-phase images are sufficient to provide adequate information for the segmentation of liver tumors.

3.6. Cross-center evaluation

To verify the generalizability of the proposed model and explore the effectiveness of the proposed method in data from different sources, we performed additional cross-center evaluation. Specifically, we split the collected data into two datasets based on their respective data sources. For training, we utilized a multi-phase contrast-enhanced CT dataset obtained from Sun Yat-Sen University's Third Affiliated Hospital (108 patients). In contrast, for testing, we employed a dataset from a distinct center, Sun Yat-Sen University's Fifth Affiliated Hospital (56 patients). Table 6 presents a detailed comparison of the segmentation performance between our method and the current state-of-the-art approach. We observed that the multi-phase methods achieved an improved performance compared to the single-phase methods. Our method has the best DSC (83.64%), Sensitivity (93.45%), ARVD (7.87%), and ASSD (5.19 mm) compared to the state-of-the-art methods.

Table 6. Performance comparison $(MeanValue \pm StandardDeviation)$ of different models on the cross-center dataset (Sun Yat-Sen University's Third Affiliated Hospital and Sun Yat-Sen University's Fifth Affiliated Hospital).

Methods	DSC (%)	Sensitivity (%)	ARVD (%)	ASSD (mm)
Single-phase				
U-Net	76.61 ± 6.28	89.20 ± 5.54	8.24 ± 2.61	6.51 ± 2.37
TransUnet	80.60 ± 5.97	91.73 ± 5.41	8.24 ± 3.09	5.67 ± 1.76
SegFormer	81.79 ± 5.82	92.30 ± 5.07	8.05 ± 2.98	5.07 ± 1.66
Multi-phase				
MW-UNet	81.92 ± 6.06	92.75 ± 4.86	8.22 ± 3.07	5.58 ± 1.88
PA-ResSeg	82.14 ± 6.06	92.84 ± 4.86	8.22 ± 3.06	5.51 ± 1.87
Ours	83.64 ± 6.16	93.45 ± 4.93	7.87 ± 2.58	5.19 ± 1.98

Note: *T-test was used to examine the differences in performance between the state-of-the-art methods and our method (Ours vs U-Net, Ours vs TransUnet, Ours vs SegFormer, Ours vs MW-UNet, Ours vs PA-ResSeg). P-values of DSC are all $P < 0.05$; P-values of Sensitivity are all $P < 0.05$; P-values of ARVD are $P = 0.26$, $P < 0.05$, $P < 0.05$, $P < 0.05$, and $P < 0.05$; P-values of ASSD are $P = 0.10$, $P = 0.18$, $P = 0.46$, $P < 0.05$, and $P < 0.05$.

3.7. Application in MVI prediction

In this investigation, we evaluated MI-TransSeg's clinical applicability by leveraging tumor masks obtained via various multi-phase segmentation methods to assess the microvascular invasion (MVI) in hepatocellular carcinoma (HCC). Supporting evidence from radiomics research demonstrated that the predictive performance of deep learning models in MVI prediction, particularly when utilizing radiomic features from high-quality segmentation, was effective in predicting MVI [60,61]. The utility of specific radiomic signatures in forecasting MVI has been documented, with potential applications in developing both predictive and prognostic models [62]. Central to this endeavor is the accuracy of tumor segmentation, as it directly affects the integrity of radiomic features, and consequently, the predictive validity of MVI models [63]. This interrelation prompted us to meticulously extract radiomic features to classify the MVI, adhering to a rigorously defined protocol.

First, we utilized an open-source Python package (Pyradiomics) [64] to extract the radiomic features of the liver tumor, which were delineated by different liver segmentation methods. In this step, 107 radiomic features were obtained from 14 shape features, 18 first-order features, 24 gray level co-occurrence matrix (GLCM) features, 16 gray level run-length matrix (GLRLM) features, 16 gray level size region matrix (GLSZM) features, 14 gray level dependence matrix (GLDM) features, and 5 neighborhood gray-tone difference matrix (NGTDM) features. The more accurate the tumor segmentation, the more precise radiomic features we can obtain. Next, we applied two fully connected layers and nonlinear activation functions to predict the results of the MVI.

The MVI prediction accuracy using different tumors masks obtained by different segmentation methods was evaluated by a receiver operating characteristic (ROC) curve and the area under the curve (AUC), as shown in Figure 7. It can be seen that our proposed MI-TransSeg had the best AUC among the methods, which indicates that the liver tumor mask segmented by our method has a greater potential to be applied for clinical application of MVI prediction.

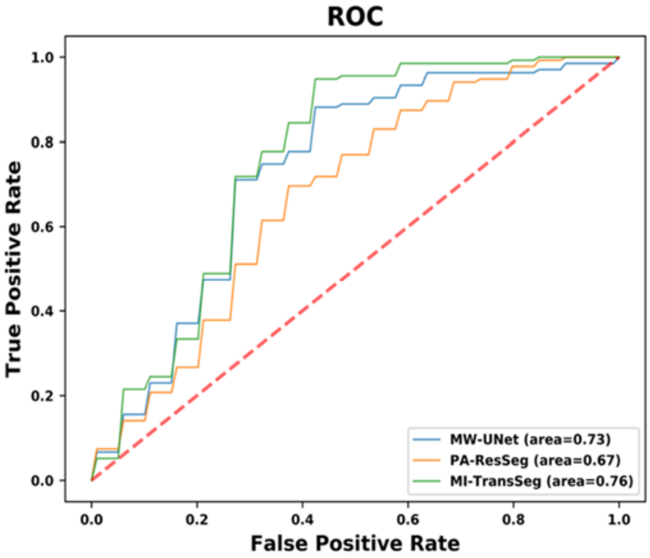

Figure 7. Receiver operating characteristic curves (ROCs) of three methods (MW-Unet, PA-ResSeg and MI-TransSeg) and their corresponding AUC values.

4. Discussion

The successful automated segmentation of liver tumors is vital for clinical decision-making and the treatment of liver-related conditions. The segmentation of liver tumors poses significant challenges for automation due to the unclear borders between nearby organs and a minimal contrast in intensity. To address this challenge, our research proposed an innovative technique to segment liver tumors across multiple phases. Results from our extensive experimentation on clinical multi-phase contrast-enhanced CT liver tumor datasets showed that our method is superior to others.

Our MI-TransSeg network architecture offers several distinctive advantages for liver tumor segmentation tasks. It incorporates multi-resolution scales feature aggregation strategy, thus enhancing the utility of multi-phase data, which we have found to be particularly effective when aggregating features across two resolution scales. The hierarchical design of our network capitalizes on both low- and high-resolution features, thus culminating in superior segmentation outcomes. Furthermore, the integration of the Transformer's self-attention mechanism broadens the network's receptive field, thus enriching its feature representation capabilities. Additionally, the introduction of a PV feature token within the decoder has been instrumental in refining the tumor boundary delineation. Crucially, our network's multi-phase feature interaction module optimizes the synergy of phase-specific information, which minimizes false positives and improves the accuracy of segmentation. We believe that these findings will catalyze further exploration into machine learning methodologies for multi-phase CT analyses in liver tumors, potentially revolutionizing the assessment of microvascular invasion and aiding in the preoperative strategizing of surgical interventions.

For the choice of the number of aggregated resolution scales, the choice will vary under different tasks. For tasks that need to obtain more refined features, such as the segmentation of liver tumors, we need to aggregate features at multiple resolution scales to utilize more useful information. However, aggregating features at more resolution levels, such as MSFA-4, cannot further improve the performance of tumor segmentation, as too much information may lead to problems such as blurred tumor boundaries. For other tasks, aggregating features at four resolution scales may be effective to improve the accuracy of segmentation.

As to the choice of the number of phases to use, it also needs to be decided according to the actual task. For the task of multi-phase liver tumors segmentation, we observed that the use of two phases is already sufficient to provide adequate information for liver tumor segmentation, while the use of more phases cannot effectively improve the accuracy of tumor segmentation.

While our method offers numerous advantages, it is important to acknowledge certain limitations. One such limitation pertains to clinical multi-phase CT data, where in relatively rare cases, the parameters of the scan protocol may be adjusted for contrast-enhanced phases for better observation in clinic. Such adjustments may result in relatively significant changes in the alignment of the start slice and end slice positions among different phases, ultimately impacting the accuracy of multi-phase liver tumor segmentation methods. Another limitation is the relatively limited training data available from two hospitals. To bolster the reliability and applicability of our model, we are committed to expanding our dataset. Future research endeavors will involve the inclusion of data from additional hospitals, thereby enhancing the generalizability of our model and ensuring its effectiveness across diverse healthcare settings.

5. Conclusions

In conclusion, our Transformer-based multi-phase liver tumor segmentation network shows a superior segmentation performance, with the potential to enhance current segmentation techniques. Our network incorporates a multi-phase feature aggregator strategy, hierarchical structure, and PV feature token, which collectively enable the network to effectively reduce the occurrence of false segmentation and generate highly refined and well-profiled segmentation results. Our study demonstrates the significant benefits of utilizing multi-phase images, which provides rich and valuable information for tumor segmentation and greatly enhances the segmentation accuracy. Overall, the proposed network represents a significant step forward in the field of medical image segmentation and holds great potential for future clinical applications.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the National Science Foundation of China under grant 81801809, Guangdong Basic and Applied Basic Research Foundation under grant 2023A1515011488 and 2020A1515010572, and Zhuhai Basic and Applied Basic Research Foundation under Grant ZH22017003200001PWC. This research was also funded by Shantou University (STU Scientific Research Foundation for Talents: NTF21004).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.*, **71** (2021), 209–249. https://doi.org/10.3322/caac.21660
2. J. M. Llovet, R. K. Kelley, A. Villanueva, A. G. Singal, E. Pikarsky, S. Roayaie, et al., Hepatocellular carcinoma, *Nat. Rev. Dis. Primers*, **7** (2021), 6. https://doi.org/10.1038/s41572-020-00240-3
3. F. X. Bosch, J. Ribes, M. Díaz, R. Cléries, Primary liver cancer: worldwide incidence and trends, *Gastroenterology*, **127** (2004), S5–S16. https://doi.org/10.1053/j.gastro.2004.09.011
4. X. Wu, J. Li, C. Wang, G. Zhang, N. Zheng, X. Wang, Application of different imaging methods in the early diagnosis of primary hepatic carcinoma, *Gastroenterol. Res. Pract.*, **2016** (2016), 8763205. https://doi.org/10.1155/2016/8763205
5. K. Song, D. Wu, Shared decision-making in the management of patients with inflammatory bowel disease, *World J. Gastroenterol.*, **28** (2022), 3092–3100. https://doi.org/10.3748%2Fwjg.v28.i26.3092

6. C. Chang, H. Chen, Y. Chang, M. Yang, C. Lo, W. Ko, et al., Computer-aided diagnosis of liver tumors on computed tomography images, *Comput. Methods Programs Biomed.*, **145** (2017), 45–51. https://doi.org/10.1016/j.cmpb.2017.04.008
7. W. Li, F. Jia, Q. Hu, Automatic segmentation of liver tumor in CT images with deep convolutional neural networks, *J. Comput. Commun.*, **3** (2015), 146–151. http://dx.doi.org/10.4236/jcc.2015.311023
8. R. Naseem, Z. A. Khan, N. Satpute, A. Beghdadi, F. A. Cheikh, J. Olivares, Cross-modality guided contrast enhancement for improved liver tumor image segmentation, *IEEE Access*, **9** (2021), 118154–118167. https://doi.org/10.1109/ACCESS.2021.3107473
9. L. Wang, M. Wu, R. Li, X. Xu, C. Zhu, X. Feng, MVI-Mind: A novel deep-learning strategy using computed tomography (CT)-based radiomics for end-to-end high efficiency prediction of microvascular invasion in hepatocellular carcinoma, *Cancers*, **14** (2022), 2956. https://doi.org/10.3390/cancers14122956
10. Y. Jiang, S. Cao, S. Cao, J. Chen, G. Wang, W. Shi, et al., Preoperative identification of microvascular invasion in hepatocellular carcinoma by XGBoost and deep learning, *J. Cancer Res. Clin. Oncol.*, **147** (2021), 821–833. https://doi.org/10.1007/s00432-020-03366-9
11. A. Radtke, S. Nadalin, G. C. Sotiropoulos, E. P. Molmenti, T. Schroeder, C. Valentin-Gamazo, et al., Computer-assisted operative planning in adult living donor liver transplantation: A new way to resolve the dilemma of the middle hepatic vein, *World J. Surg.*, **31** (2007), 175–185. https://doi.org/10.1007/s00268-005-0718-1
12. P. Liang, Y. Wang, X. Yu, B. Dong, Malignant liver tumors: treatment with percutaneous microwave ablation—complications among cohort of 1136 patients, *Radiology*, **251** (2009), 933–940. https://doi.org/10.1148/radiol.2513081740
13. S. Gul, M. S. Khan, A. Bibi, A. Khandakar, M. A. Ayari, M. E. H. Chowdhury, Deep learning techniques for liver and liver tumor segmentation: A review, *Comput. Biol. Med.*, **147** (2022), 105620. https://doi.org/10.1016/j.compbiomed.2022.105620
14. L. Soler, H. Delingette, G. Malandain, J. Montagnat, N. Ayache, C. Koehl, et al., Fully automatic anatomical, pathological, and functional segmentation from CT scans for hepatic surgery, *Comput. Aided Surg.*, **6** (2001), 131–142. https://doi.org/10.3109/10929080109145999
15. H. A. Nugroho, D. Ihtatho, H. Nugroho, Contrast enhancement for liver tumor identification, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **41** (2008), 201. https://doi.org/10.54294/1uhwld
16. M. Esfandiarkhani, A. H. Foruzan, A generalized active shape model for segmentation of liver in low-contrast CT volumes, *Comput. Biol. Med.*, **82** (2017), 59–70. https://doi.org/10.1016/j.compbiomed.2017.01.009
17. D. Wong, J. Liu, F. Yin, Q. Tian, W. Xiong, J. Zhou, et al., A semi-automated method for liver tumor segmentation based on 2D region growing with knowledge-based constraints, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **41** (2008), 159. https://doi.org/10.54294/25etax
18. L. Fernandez-de-Manuel, J. L. Rubio, M. J. Ledesma-Carbayo, J. Pascau, J. M. Tellado, E. Ramon, et al., 3D liver segmentation in preoperative CT images using a level-sets active surface method, in *International Conference of the IEEE Engineering in Medicine and Biology Society*, (2009), 3625–3628. https://doi.org/10.1109/iembs.2009.5333760

19. S. S. Kumar, R. S. Moni, J. Rajeesh, An automatic computer-aided diagnosis system for liver tumours on computed tomography images, *Comput. Electr. Eng.*, **39** (2013), 1516–1526. https://doi.org/10.1016/j.compeleceng.2013.02.008
20. R. Kaur, L. Kaur, S. Gupta, Enhanced K-mean clustering algorithm for liver image segmentation to extract cyst region, in *IJCA Special Issue on Novel Aspects of Digital Imaging Applications*, **1** (2011), 59–66.
21. T. Zhou, S. Canu, S. Ruan, Fusion based on attention mechanism and context constraint for multi-modal brain tumor segmentation, *Comput. Med. Imaging Graphics*, **86** (2020), 101811. https://doi.org/10.1016/j.compmedimag.2020.101811
22. J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, I. B. Ayed, HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation, *IEEE Trans. Med. Imaging*, **38** (2018), 1116–1126. https://doi.org/10.1109/TMI.2018.2878669
23. Q. Yu, Y. Shi, J. Sun, Y. Gao, J. Zhu, Y. Dai, Crossbar-net: a novel convolutional neural network for kidney tumor segmentation in CT images, *IEEE Trans. Image Process.*, **28** (2019), 4060–4074. https://doi.org/10.1109/TIP.2019.2905537
24. X. Ma, L. M. Hadjiiski, J. Wei, H. P. Chan, K. H. Cha, R. H. Cohan, et al., U-Net based deep learning bladder segmentation in CT urography, *Med. Phys.*, **46** (2019), 1752–1765. https://doi.org/10.1002/mp.13438
25. P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, et al., Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2016), 415–423. https://doi.org/10.1007/978-3-319-46723-8_48
26. G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, H. Meine, Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing, *Sci. Rep.*, **8** (2018), 15497. https://doi.org/10.1038/s41598-018-33860-7
27. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
28. C. Li, Y. Tan, W. Chen, X. Luo, Y. Gao, X. Jia, et al., Attention Unet++: A nested attention-aware U-Net for liver CT image segmentation, in *IEEE International Conference on Image Processing*, (2020), 345–349. https://doi.org/10.1109/ICIP40778.2020.9190761
29. H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, et al., Unet 3+: A full-scale connected unet for medical image segmentation, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (2020), 1055–1059. https://doi.org/10.1109/ICASSP40776.2020.9053405
30. H. Seo, C. Huang, M. Bassenne, R. Xiao, L. Xing, Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images, *IEEE Trans. Med. Imaging*, **39** (2019), 1316–1325. https://doi.org/10.1109/TMI.2019.2948320
31. D. T. Kushnure, S. N. Talbar, MS-UNet: A multi-scale UNet with feature recalibration approach for automatic liver and tumor segmentation in CT images, *Comput. Med. Imaging Graphics*, **89** (2021), 101885. https://doi.org/10.1016/j.compmedimag.2021.101885

32. X. Xu, Q. Zhu, H. Ying, J. Li, X. Cai, S. Li, et al., A knowledge-guided framework for fine-grained classification of liver lesions based on multi-phase CT images, *IEEE J. Biomed. Health Inf.*, **27** (2023), 386–396. https://doi.org/10.1109/JBHI.2022.3220788
33. W. Shi, S. Kuang, S. Cao, B. Hu, S. Xie, S. Chen, et al., Deep learning assisted differentiation of hepatocellular carcinoma from focal liver lesions: choice of four-phase and three-phase CT imaging protocol, *Abdom. Radiol.*, **45** (2020), 2688–2697. https://doi.org/10.1007/s00261-020-02485-8
34. Y. Xu, M. Cai, L. Lin, Y. Zhang, H. Hu, Z. Peng, et al., PA-ResSeg: A phase attention residual network for liver tumor segmentation from multiphase CT images, *Med. Phys.*, **48** (2021), 3752–3766. https://doi.org/10.1002/mp.14922
35. I. R. Kamel, M. A. Choti, K. M. Horton, H. J. V. Braga, B. A. Birnbaum, E. K. Fishman, et al., Surgically staged focal liver lesions: accuracy and reproducibility of dual-phase helical CT for detection and characterization, *Radiology*, **227** (2003), 752–757. https://doi.org/10.1148/radiol.2273011768
36. F. Ouhmich, V. Agnus, V. Noblet, F. Heitz, P. Pessaux, Liver tissue segmentation in multiphase CT scans using cascaded convolutional neural networks, *Int. J. Comput. Assisted Radiol. Surg.*, **14** (2019), 1275–1284. https://doi.org/10.1007/s11548-019-01989-z
37. C. Sun, S. Guo, H. Zhang, J. Li, M. Chen, S. Ma, et al., Automatic segmentation of liver tumors from multiphase contrast-enhanced CT images based on FCNs, *Artif. Intell. Med.*, **83** (2017), 58–66. https://doi.org/10.1016/j.artmed.2017.03.008
38. Y. Wu, Q. Zhou, H. Hu, G. Rong, Y. Li, S. Wang, Hepatic lesion segmentation by combining plain and contrast-enhanced CT images with modality weighted U-Net, in *IEEE International Conference on Image Processing*, (2019), 255–259. https://doi.org/10.1109/ICIP.2019.8802942
39. Y. Zhang, C. Peng, L. Peng, H. Huang, R. Tong, L. Lin, et al., Multi-phase liver tumor segmentation with spatial aggregation and uncertain region inpainting, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 68–77. https://doi.org/10.1007/978-3-030-87193-2_7
40. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems*, **30** (2017).
41. L. Wang, X. Wang, B. Zhang, X. Huang, C. Bai, M. Xia, et al., Multi-scale Hierarchical Transformer structure for 3D medical image segmentation, in *IEEE International Conference on Bioinformatics and Biomedicine*, (2021), 1542–1545. https://doi.org/10.1109/BIBM52615.2021.9669799
42. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, et al., Swin-unet: Unet-like pure transformer for medical image segmentation, in *European Conference on Computer Vision*, (2021), 205–218. https://doi.org/10.1007/978-3-031-25066-8_9
43. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., Transunet: Transformers make strong encoders for medical image segmentation, preprint, arXiv:2102.04306. https://doi.org/10.48550/arXiv.2102.04306
44. H. Xiao, L. Li, Q. Liu, X. Zhu, Q. Zhang, Transformers in medical image segmentation: A review, *Biomed. Signal Process.*, **84** (2023), 104791. https://doi.org/10.1016/j.bspc.2023.104791
45. K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, et al., Transformers in medical image analysis, *Intell. Med.*, **3** (2023), 59–78. https://doi.org/10.1016/j.imed.2022.07.002

46. Y. Xu, X. He, G. Xu, G. Qi, K. Yu, L. Yin, et al., A medical image segmentation method based on multi-dimensional statistical features, *Front. Neurosci.*, **16** (2022), 1009581. https://doi.org/10.3389/fnins.2022.1009581
47. X. He, G. Qi, Z. Zhu, Y. Li, B. Cong, L. Bai, Medical image segmentation method based on multi-feature interaction and fusion over cloud computing, *Simul. Modell. Pract. Theory*, **126** (2023), 102769. https://doi.org/10.1016/j.simpat.2023.102769
48. A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 272–284. https://doi.org/10.1007/978-3-031-08999-2_22
49. Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, *Inf. Fusion*, **91** (2023), 376–387. https://doi.org/10.1016/j.inffus.2022.10.022
50. Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, Y. Liu, X-Net: a dual encoding–decoding method in medical image segmentation, *Visual Comput.*, **39** (2023), 2223–2233. https://doi.org/10.1007/s00371-021-02328-7
51. J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, Medical Transformer: Gated axial-attention for medical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 36–46. https://doi.org/10.1007/978-3-030-87193-2_4
52. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, in *Advances in Neural Information Processing Systems*, **34** (2021), 12077–12090.
53. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in *International Conference on Learning Representations*, preprint, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
54. C. Peng, Y. Zhang, J. Zheng, B. Li, J. Shen, M. Li, et al., IMIIN: an inter-modality information interaction network for 3D multi-modal breast tumor segmentation, *Comput. Med. Imaging Graphics*, **95** (2022), 102021. https://doi.org/10.1016/j.compmedimag.2021.102021
55. L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. H. Jiang, et al., Tokens-to-token vit: Training vision transformers from scratch on imagenet, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 558–567. https://doi.org/10.48550/arXiv.2101.11986
56. N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 4722–4732.
57. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, et al., Automatic differentiation in pytorch, in *Advances in Neural Information Processing Systems*, 2017.
58. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, preprint, arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980
59. W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, **29** (2016).
60. W. Zhou, W. Jian, X. Cen, L. Zhang, H. Guo, Z. Liu, et al., Prediction of microvascular invasion of hepatocellular carcinoma based on contrast-enhanced MR and 3D convolutional neural networks, *Front. Oncol.*, **11** (2021), 588010. https://doi.org/10.3389/fonc.2021.588010

61. X. Zhong, H. Long, L. Su, R. Zheng, W. Wang, Y. Duan, et al., Radiomics models for preoperative prediction of microvascular invasion in hepatocellular carcinoma: a systematic review and meta-analysis, *Abdom. Radiol.*, **47** (2022), 2071–2088. https://doi.org/10.1007/s00261-022-03496-3

62. K. Bera, N. Braman, A. Gupta, V. Velcheti, A. Madabhushi, Predicting cancer outcomes with radiomics and artificial intelligence in radiology, *Nat. Rev. Clin. Oncol.*, **19** (2022), 132–146. https://doi.org/10.1038/s41571-021-00560-7

63. J. Liu, D. Cheng, Y. Liao, C. Luo, Q. Lei, X. Zhang, et al., Development of a magnetic resonance imaging-derived radiomics model to predict microvascular invasion in patients with hepatocellular carcinoma, *Quant. Imaging Med. Surg.*, **13** (2023), 3948–3961. https://doi.org/10.21037/qims-22-1011

64. J. J. M. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, et al., Computational radiomics system to decode the radiographic phenotype, *Cancer Res.*, **77** (2017), e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339