IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 November 2022, accepted 21 December 2022, date of publication 9 January 2023, date of current version 13 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234997

RESEARCH ARTICLE

Multi-Scale Deep Information and Adaptive Attention Mechanism Based Coronary Reconstruction of Superior Mesenteric Artery

KUN ZHANG[1,4,5], (Member, IEEE), YU HAN[1], PEIXIA XU[1], MEIRONG WANG[2], JUSHUN YANG[2], PENGCHENG LIN[1], DANNY CROOKES[3], (Senior Member, IEEE), BOSHENG HE[2,5,6], AND LIANG HUA[1]

[1]School of Electrical Engineering, Nantong University, Nantong, Jiangsu 226001, China
[2]Department of Radiology, Affiliated Hospital 2 of Nantong University, Nantong, Jiangsu 226001, China
[3]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN Belfast, U.K.
[4]Nantong Key Laboratory of Intelligent Control and Intelligent Computing, Nantong, Jiangsu 226001, China
[5]Nantong Key Laboratory of Intelligent Medicine Innovation and Transformation, Nantong, Jiangsu 226001, China
[6]Clinical Medicine Research Center, Affiliated Hospital 2 of Nantong University, Nantong, Jiangsu 226001, China

Corresponding authors: Bosheng He (boshenghe@126.com) and Liang Hua (hualiang@ntu.edu.cn)

This work was supported in part by the Ethics Committee of the Affiliated Hospital 2 of Nantong University under Grant 2020YKS024; in part by the Nantong Health Commission, in 2020, under Grant QA2020002; in part by the Key Clinical Project of Nantong University under Grant 2019LZ002; and in part by the Key Scientific Grant from Jiangsu Commission of Health under Grant ZD2021059.

ABSTRACT Vascular images contain a lot of key information, such as length, diameter and distribution. Thus reconstruction of vessels such as the Superior Mesenteric Artery is critical for the diagnosis of some abdominal diseases. However automatic segmentation of abdominal vessels is extremely challenging due to the multi-scale nature of vessels, boundary-blurring, low contrast, artifact disturbance and vascular cracks in Maximum Intensity Projection images. In this work, we propose a dual attention guided method where an adaptive adjustment field is applied to deal with multi-scale vessel information, and a channel feature fusion module is used to refine the extraction of thin vessels, reducing the interference and background noise. In particular, we propose a novel structure that accepts multiple sequential images as input, and successfully introduces spatial-temporal features by contextual information. A further IterUnet step is introduced to connect tiny cracks caused using CT scans. Comparing our proposed model with other state-of-the-art models, our model yields better segmentation and achieves an average F1 metric of 0.812.

INDEX TERMS Superior mesenteric artery, context-guided, sequential image segmentation, multi-scale information, adaptive attention.

I. INTRODUCTION

Mesenteric vascular disease is a disease where the mesenteric arteries or veins are continuously damaged for various reasons. Among these, superior mesenteric artery embolism has an overall mortality of 60% to 80% for its insidious onset and rapid progression [1]. Such emergency cases pose high demands for accuracy and efficiency of diagnosis. Medical images contain key vessel information for diagnosis and treatment [2], [3], [4]; for example, patients with thrombosis tend to have local thickening diameter, and in Crohn's disease, the mesentery is often dilated and twisted, with abundant marginal capillaries. Accurate segmentation of the Superior Mesenteric Artery (SMA) is a critical step for the diagnosis of some abdominal vascular diseases. Doing this automatically alleviate the workload of radiologists and other medical experts in clinical scenarios.

In the angiogram obtained by a CT scan, each CT slice contains only a small proportion of target vessels, so selecting target vessels from raw CT images is extremely challenging. At present, clinical doctors obtain vascular information with the help of the Maximum Itensity Projection (MIP)

The associate editor coordinating the review of this manuscript and approving it for publication was Fahmi Khalifa

FIGURE 1. 3D angiograph generated from thick MIP. SMA originates from the Abdominal Aorta (AA), as indicated by the arrow.

technique [5], which can be divided into thin MIP and thick MIP. Thick MIP can be used to display the three-dimensional structure of vessels at the cost of overlooking small vessels, as seen in Fig. 1.

This shortcoming can be made up for by using thin MIP. Radiologists can project the voxel of the greatest CT value at a certain thickness (often 15mm) onto the background plan by extracting microvascular structures from various CT depth layers. Although features of thin vessels can be retained, no single thin MIP image can reflect the coronal plane of the vessel's maximal projection [6]. The maximum coronal projection mentioned in this paper is intended to show the main vessels of SMA while ensuring their consistency. Radiologists mentally fuse two complementary data sets for diagnosis, with the risk of leaving out substantial amounts of data.

Based on the imaging scenario of thin MIP, this paper aims to generate a complete coronal view of blood vessels based on the thin MIP sequence. To our knowledge, no existing studies have been done on SMA vascular reconstruction, while studies on MIP sequences are relatively rare. Although the MIP image filters most of the soft tissue and abdominal interference in the original CT sequence, the use of the MIP image inevitably produces a large number of arterial interferences with similar feature expressions.

In recent years, deep learning methods represented by Deep Convolutional Neural N6etworks (DCNN) [7], [8], [9] have been widely used in medical imaging tasks. They have the ability to automatically extract effective high-level features by learning a large number of labeled samples. In this scenario, the encoder-decoder structure has become the first choice in dealing with medical images. The challenges of convolutional networks for SMA vessel segmentation in MIP sequences lie in multi-scale vessel information, wide distribution of low contrast and inference of adjective vessels such as AA and vein vessels. Narrow vessels are very challenging to

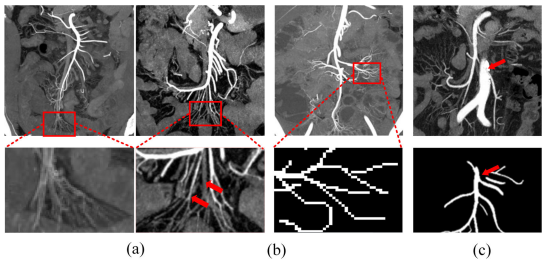


FIGURE 2. Challenges for accurate segmentation of SMA. From (a) to (c), each part displays the blurring of thin and peripheral vessels due to low contrast, tiny vessels a few pixels wide, vascular discontinuities due to imaging techniques, and occluded tissue respectively.

distinguish, as some have a width of only one or a few pixels, limiting the segmentation performance of SMA. Moreover, cracks resulting from the MIP imaging principle still need to be fixed. All of the above challenges are illustrated in Fig. 2, where multi-scale information and interference are common problems that exist in image segmentation. Existing work to address the above issues simultaneously includes [10] which utilizes dilation convolutions to detect convolutional features at multiple scales. Reference [11] proposes a Scale-Aware Pyramid Fusion block which fuses features extracted by three dilated convolutions with different dilation rates to capture different scale information. Reference [12] designs a novel Unet structure with two encoders. However, simply stacking multi-scale modules with fixed receptive fields cannot effectively capture multi-scale features and deal thoroughly with the large-scale variation of retinal vessels. It also introduces too many additional parameters. Reference [13] modifies the vessel segmentation task to a three-class classification problem between large vessels, small vessels, and background regions to reduce the problem of intra-class variation. Though these strategies may be beneficial, non-adaptive extraction methods cannot thoroughly handle multi-scale vessels.

The shortcomings in the processing of SMA reconstruct arise mainly from two aspects. The first is the large area covered by AA. To solve this issue, space contextual information can be useful since several parts of one vessel usually appear in adjacent sequences. 3D convolution is a suitable choice because it is a standard computational structure and has achieved great success in extracting spatial information. Second, the information loss caused by the thickness of the CT scan makes the continuous MIP projection plane still unable to reflect the complete vascular structure. This results in a fracture in the final superimposed 2D segmentation result. To address this, IterUnet is proposed to deduce the missing part based on the trend of vessels. Despite the development of numerous vessel segmentation algorithms in recent years, few experts have focused on finding a solution to the issue of vessel cracks brought about by insufficient original data.

In this paper, we propose a vessel reconstruction model based on the MIP sequence. The main contributions of our model include:

1) A novel backbone context-guided sequential structure (CGS) is proposed which uses 3D convolution and 3D

residual blocks in the encoder to refine spatial features of sequential images, the corresponding serial structure of 2D convolution and 2D residual blocks in the decoder is used for feature recovery. Redundant features in the space are integrated into the target vessels to obtain the maximum vessel coronal plane generated from (overlapping) sequences of 4 adjacent images.

2) A proposed dual-attention-guided structure (DA) includes a scale-aware attention module and a feature fusion module (FFM), which generates deformable convolution in the scale-aware attention module, adaptively extracts vessel features, and enhances the refinement of fine vessels. FFM is introduced to assign different weights to multi-scale features during up-sampling and down-sampling, achieving precise localization of high-resolution target vessel semantic information and enhancing the ability to identify target vessels from complex backgrounds and artifacts.

3) An IterUnet step is used to solve the problem of vessel cracks caused by information loss between slices, ensuring the integrity of the vascular topological structure.

II. RELATED WORK

The starting point of our novel reconstruction method is the recent success of the U-shape architecture. As mentioned above, artificial interference, limited space information, and multi-scale vessels can all damage a vessel's accuracy and integrity. In the following part, we briefly review related methods, dividing them into two sub-areas, *i.e.* the blood vessel segmentation and multi-scale feature extraction.

A. VESSEL SEGMENTATION

In recent decades, many algorithms for vessel segmentation have emerged. Some of the most successful are based on pairwise classifiers, such as Random Forests [14]. Reference [15] presents a generative model for image synthesis that yields a probabilistic segmentation of abnormalities. Reference [16] incorporates vessel structure segmentation results obtained from a multi-atlas label propagation approach to provide strong tissue-class priors to Random Forests.

Compared with traditional methods, deep learning performs better in terms of inference, speed and generalization capacity. In the field of medical image processing, Convolution Neural Networks (CNN), Full Convolution Networks (FCN) [17], [18], [19], Recurrent Neural Networks (RNN) [20], [21], Resnet [22], Unet, and many of their variants have become mainstream.

B. MULTI-SCALE FEATURE EXTRACTION

Although some classical networks have achieved impressive performance, their ability to extract multi-scale features under low contrast is still inadequate. How to extract multi-scale features adaptively is a hot topic at present. To accomplish this goal, [23] captures target objects at different scales with the help of a dynamic kernel selection mechanism, while [24] applies a coarse-to-fine depth architecture to learn multi-scale features of multimodal images. Reference [11] proposes a Scale-Aware Pyramid Fusion block that fuses features extracted by three dilated convolutions with different dilation rates to capture different scale information. However, conducting convolutions on a large number of high-resolution feature maps requires significant GPU resources [25]. Reference [12] proposes a novel Unet structure with two encoders, but simply stacking multi-scale modules with fixed receptive fields [26] cannot effectively capture multi-scale features of SMA vessels. It also introduces too many additional parameters, and non-adaptive feature extraction prevents efficient handling of complicated situations including multi-scale information and vessels that appear to cross over. In [27], deformable convolution and deformable ROI pooling are used as substitutes to replace the original feature extraction structure, achieving the capability of dynamically adjusting the receiving field based on the input image. Prompted by this work, we designed a scale-aware attention module, which dynamically adjust the receptive field in attention to refine the feature map and extract multi-scale information.

III. METHODS

The purpose of this work is to establish a novel framework to obtain the main branches of the SMA and vessel trend. We propose a novel network architecture CGS that implements a mapping of one output corresponding to multiple inputs, integrating redundant features in space into the target vessel to obtain the maximum coronal vessel generated by 4 adjacent sequences. We then propose two sub-modules that can be embedded in CGS regarding DA and IterUnet. We first introduce DA which is consistent with the scale-aware attention module and FFM. The former dynamically adjusts the field to efficiently extract multi-scale features, and the latter fuses features from the up-sampling layers and skip layers, alleviating the interference of background artifacts while enhancing the performance of segmenting thin vessels. Finally, we apply IterUnet to fix cracks that arise from the information loss between sections, guaranteeing the connectivity of vessels. The overall architecture of the system can be seen in Fig. 3 and the pseudocode of whole network is described in Algorithm 1. Fig. 3 can be divided into two parts. On the left is the CGS architecture used to obtain the initial results of blood vessels. The purple and pink modules in the figure are two sub-modules of DA. On the right is IterUnet, which repairs the original results. Next, we will explain our structural details from the network architecture, DA, and IterUnet respectively.

A. NETWORK ARCHITECTURE

A MIP sequence is different from the general original sequence. Each MIP image is a superposition result of vessels of a certain thickness after processing by maximum density projection technology. From the thin MIP sequence, we can obtain the general trend of SMA which can be presented in the 2D plane. From the perspective of the input images, we take 4 adjacent slices as input (3D input); the superposition results

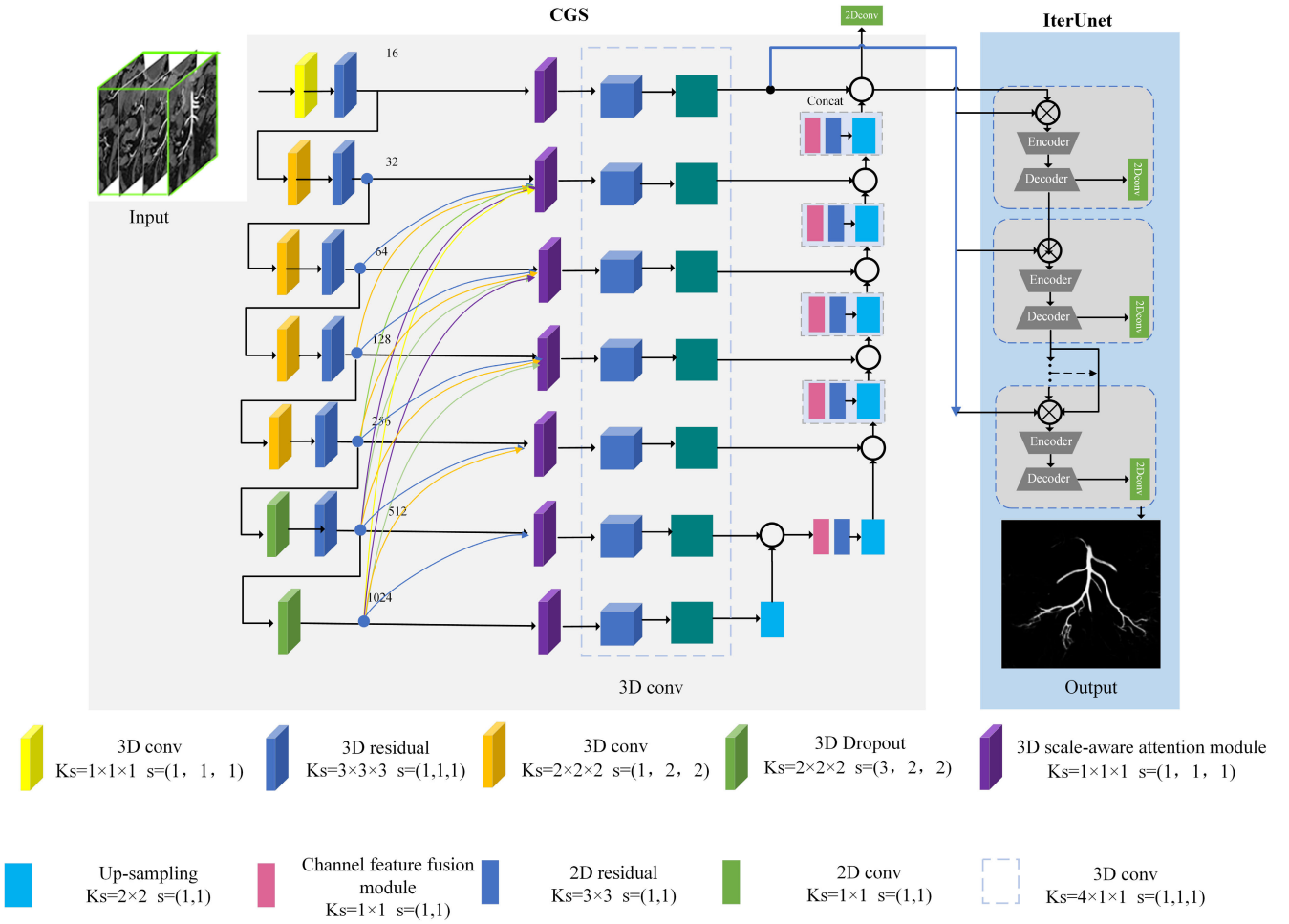


FIGURE 3. **Overview of the structure of the proposed complete system with its CGS backbone network.**

of these adjacent slices are taken as the output (2D output). In the subsequent experiment, we will provide the experimental rationale for using 4 slices as the input. As shown in Fig. 4, every four consecutive MIP images generate one segmentation result. We define such a 4 to 1 mapping as a single group, and the mapping of a MIP sequence to the final result is denoted as an ensemble group. In the reconstruction of the SMA, the appropriate sequence length was selected according to the degree of vascular extension in the coronal plane: when the patient's vascular extension is small, a length of around 15 is sufficient to reflect the trend of major vessels. But for SMA with large extensions or to get more information about marginal vessels, the length required is 30 or more. The exact number of valid slices depends on the thickness of the patient's vessels.

Unlike previous work such as 2D Unet and 3D Unet, our proposed CGS is neither a pure two-dimensional segmentation nor a three-dimensional segmentation, but a hybrid 3D-2D model that uses 3D convolution and 2D convolution alternately. It has been shown that 2D Unet performs well in two-dimensional segmentation tasks. A typical example is the segmentation of the retina. However, directly applying a 2D segmentation algorithm to the MIP sequence will overlook the relationships between slices, and key spatial context information will be ignored if it is segmented individually. In addition, networks with 3D kernels usually contain many parameters which have a high probability of gradient vanishing or exploding as well as the risk of over-fitting. Thus our method aims to combine the benefits of both the 2D and 3D approaches.

The proposed CGS network architecture can be seen in Fig.3. We make appropriate modifications based on Unet [28], which has been widely used for biomedical images since it was first proposed in 2015 and has inspired many follow-up methods. In order to address the issues mentioned above, firstly we use the series structures of convolution and residual blocks to replace the native feature extraction structure with conv+conv+pooling. The shortcut in the residual network makes it easier to optimize than the common network. This also reduces the degradation caused by the increased network depth. Secondly, we use 3D and 2D feature extractors in the encoding and decoding respectively, retaining contextual information through 3D features and reducing computation by using 2D kernels.

Algorithm 1 Pseudocode of the Proposed Network

Require:
CGS-single group consisting of 4 consecutive MIP images;
IterUnet-Input:x_0 - penultimate feature in initial network CSG, x_s^1 - feature in skip layers in CSG;

for each epoch not met **do**
  // CGS
  Convert image to feature f using 3D residual and 3D conv ;
  Fuse f comes from different stage using eq(2);
  Generate feature map y using eq(4-5);
  Obtain self-attention map using eq(7-8);
  Realize the feature dimension transformation from 3D to 2D using eq(1);
  Fuse high and low features using FFM in eq(12);
  Obtain the initial segmentation result;
  while initial segmentation result is not none **do**
    // IterUnet
    Update input by concatenating all feature maps generated by all preceding layers;
    Gain output using eq(14);
    if iterations is not met **then**
      Continue with step 10,11;
    else
      Obtain the final output;
    end if
  end while
end for

Unet uses the encoder-decoder structure, a classic architecture including a compression path and expansion path. Medical images are much more difficult to deal with than general images due to the wide distribution of gray level values and unclear boundaries. Thus, the good performance of Unet in medical images also benefits from its special processing of low-stage and high-stage features. Low-resolution information from down-sampling is good for the identification of pixel categories, and high-resolution information from up-sampling provides accurate location information for segmentation. These advantages will be inherited and preserved in our backbone.

In the first 6 layers of the CSG, 3D features $X \in \mathbb{R}^{C\times S\times H\times W}$ are extracted using the serial structure of 3D convolution and 3D residual blocks, here H, W, S and C represent height, weight, depth and channel of feature X; these features are subsequently transferred to the next layers. For grayscale images, $C = 1$, S refers to the number of slices in a single group; in this work S is taken to be 4. (In subsequent experiments, we will demonstrate the reasonableness of taking S to be 4). To prevent CGS from overfitting, a spatial pooling layer with a probability of 0.5 to discard features is embedded in layers 6 and 7, specifically placed before the convolutional layer in layer 6, while layer 7 contains only the pooling layer. The features are recovered using 2D convolution and 2D residual blocks in the decoder, and [29] states that two superimposed convolution layers in a single residual block is the optimal structure. Thus this structure is used for both 3D and 2D residual blocks in CGS (shown in Fig. 5). After the last 2D residual block, we used 2×1 convolution followed by a Sigmoid activation function to produce the final vessel mask.

To realize the feature dimension transformation, a feature dimension transformation module is designed in the skip connection layer to realize the transformation from the 3D feature to the 2D feature. Given 4 consecutive slices as input, we will achieve a set of 3D features $F' \in \mathbb{R}^{1\times 4\times 512\times 512}$. This three-dimensional feature can be decomposed into x and y along the two-dimensional plane and z along the transverse section. Hence, we realize this dimensionality reduction via the 3D convolutional kernel ($4\times1\times1$), where the first dimension of the convolutional kernel indicates the transverse axis. the paddings in the x, y, z axes are 1, 1, 0, respectively, and the strides are all 1. This operation can be illustrated in Eq (1).

$$g(x, y) = \sum_{T=-N}^{N} \sum_{W=-1}^{1} \sum_{H=-1} w(\eta, w, h) f(i + \eta, x + w, y + h) \tag{1}$$

B. DUAL ATTENTION(DA) GUIDED MODULE

With the development of deep learning, multi-scale features have shown superior performance in recent work [30], [31]. Inspired by these studies, we use attention mechanisms for adaptive feature extraction. We thus propose a dual-attention-guided adaptive channel feature fusion module that guides accurate vascular segmentation. We next discuss two components of this module in detail.

1) SCALE-AWARE ATTENTION MODULE

Recent work has proved the effectiveness of the attention mechanism [32], [33] for its superior performance in image recognition. Convolutional feature extraction ignores global information, while the attention mechanism integrates contextual information as well, which has been shown to be beneficial for improving the accuracy of image segmentation.

Methods to gain global information, for example CBAM and self-attention mechanism, compute the feature similarities between a position and all positions to capture long-range dependencies. This operation means that any two positions with similar features contribute to mutual improvement regardless of their distance in the spatial dimension. Attention-based algorithms were first proposed in natural language processing tasks where associations between words are crucial for semantic comprehension. However, for local SMA in the MIP image, we argue that not all pixels can contribute to the semantic representation, and some are even harmful. Many attemps have been made to capture more global information, among which the most popular is pyramid feature extraction. Reference [29] proposes a Scale-Aware

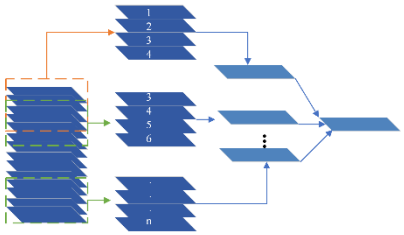


FIGURE 4. Structure of CGS.

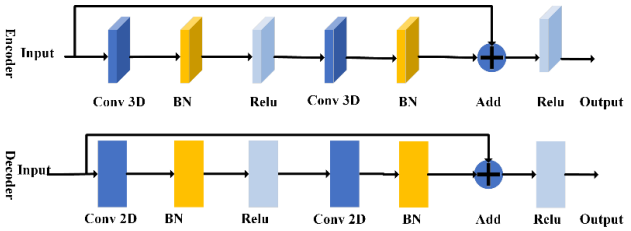


FIGURE 5. Structure of feature extractor.

Pyramid Fusion block which applies three dilated convolutions to capture different scale information. A novel Unet structure is suggested by [31] to accept features from multiple sensory fields. Regrettably, these fields depend on manual settings while the region of interest usually depends on the characteristics of the image itself. Simply stacking multi-scale features with manually fixed receptive fields fails to effectively capture multi-scale features, and cannot deal thoroughly with the large-scale variation of vessels.

One feasible approach is to capture the relationship between pixels within a reasonable area and suppress noise in the background. For this, we propose deformable convolution [34] to adaptively extract vessel features and strengthen the ability to refine thin vessels. This approach has achieved success in several tasks [35], [36], [37], [38], enabling receptive fields to automatically adapt to different sizes of vessels.

In the process of encoding, features may be progressively weakened when they are gradually transmitted to shallower layers. To better utilize the multi-scale information, starting from the last layer, we upsample features of each layer to the same size as the target layer. Feature fusion is carried out using the method shown in Fig 6. (a)(b). Taking stage 5 as an example, feature maps in stages 6,7 will be upsampled by factors of 2 and 4. Then the weight matrix is used to realize the pairwise fusion of features. The results of this is then input into the Scale-aware attention module to achieve adaptive feature extraction. The process of multi-scale fusion can be expressed in Eq(2).

$$F_{in} = \begin{cases} f_{in}^{k} = f_1, k = 1 \\ f_{in}^{k} = \underset{i=k}{\overset{7}{C}} \left(f_i \circ 2^{i-k}\right), k = \{2, \ldots, 7\} \end{cases} \tag{2}$$

where f represents the feature of each stage after 3D sampling, C refers to the operation concatenate operation, f_{in} represents the fused feature comimg from the multi-scale feature, and $\circ$ represents the upsampling operation.

A traditional convolution applies a regular grid G (e.g. 3×3 or 5×5) to sample the input feature x, utilizing some function to calculate sample values weighted by a kernel w. Supposing a convolutional kernel size with $K = t^2$ sampling locations, for each pixel p_0 in the output feature map y, the convolution process can be described in eq(3).

$$y(p_0) = \sum_{k=1}^{t^2} w_k \cdot x(p_0 + p_k) \tag{3}$$

where p_k,w_k refer to offset positions within the grid G and the corresponding kernel weight. $x()$ are sampling positions in x, t represents width (or length) of G. In a deformable convolution, G is irregular, and the shape changes according to the structural features of the input data; thus an extra parameter Δp_k is needed to record the offset of each position in G. For each pixel p_0 in the output feature map y, the deformable convolution process can be described in Eq(4).

$$y(p_0) = \sum_{k=1}^{t^2} w_k x(p_0 + p_k \Delta p_k) \tag{4}$$

We first denote $F_{in} \in \mathbb{R}^{C\times S\times H\times W}$ as the input feature, where H, W, S and C represent height, weight, depth and channel of feature F_{in}; then we feed it into 2 parallel 3D deformable convolution layers and obtain two new features F_{o1}, $F_{o2} \in \mathbb{R}^{C\times S\times H\times W}$, the transformation process is illustrated in Fig. 7.

Eq(3) is also applicable when extending deformable convolution from 2D to 3D. Here, the sample grid G will have t^3 sampling positions. The input feature F_{in} is composed of a set of voxels $\{p_i | i = 1, 2, \ldots, N\}$. The voxel at each position can be denoted as (x, y, z). The corresponding offset can be learned by an additional convolution layer, which can be described as $\Delta p_k = \{\Delta x_k, \Delta y_k, \Delta z_k\}$. We let $p = p_0 + p_k + \Delta p_k$, where p represents the sampling locations in F_{in} after adding of the offset. Since Δp_k is usually fractional, the output pixel position may not have the corresponding position in the input image, so trilinear interpolation is applied to obtain the final output, which can be found in Eq(5-6).

$$x\left(p_i\right) = \sum_{q_j \in Q} x\left(q_j\right) \cdot g\left(q^{-x}, p^{x}\right) \cdot g\left(q^{-y}, p^{y}\right) \cdot g\left(q^{-z}, p^{z}\right) \tag{5}$$

$$g(i, j) = \max(0, 1 - |i - j|) \tag{6}$$

where Q represents the set of all integer positions in the sampled volume centered on p_i. Eq(6) indicates that in trilinear interpolation points with a distance less than 1 near the target point are included in the calculation.

We next introduce a $1\times1\times1$ 3D conv to reshape F_{o1} and F_{o2} to 1/4 of original channel. After this step, the feature dimension becomes F_{o1},$F_{o2} \in \mathbb{R}^{N\times\frac{C}{4}}$. Here, $N = S \times H \times W$ means the total number of voxels. This operation helps alleviate the computation cost. We further

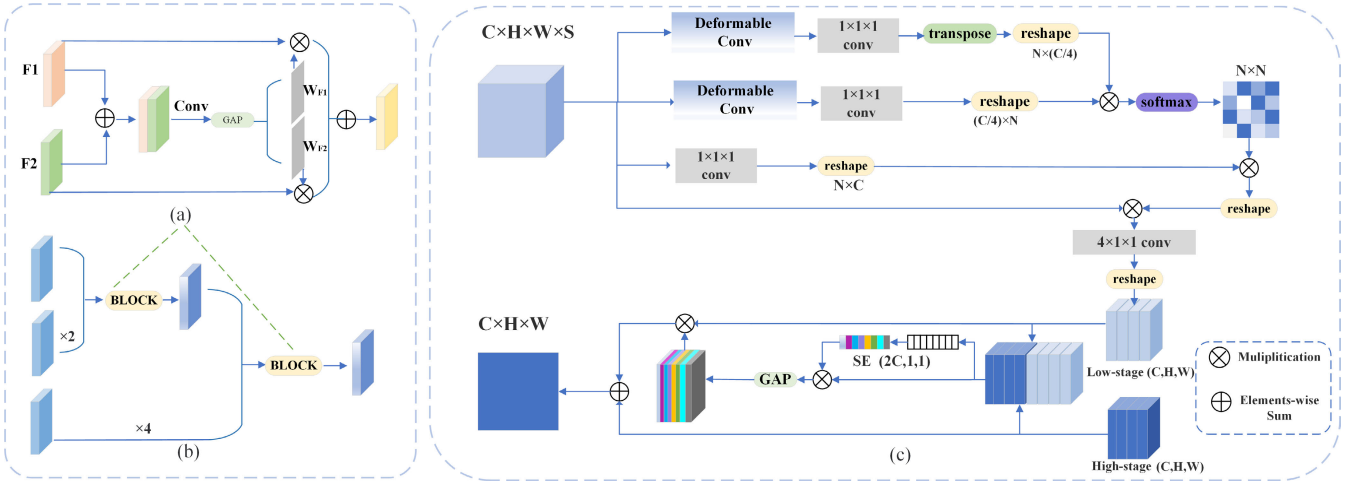


FIGURE 6. Illustration of dual attention, in which (b) can be used to obtain multi-scale information and is used as the input to the dual attention module, (a) shows details of the block in fig(b), (c) describes the overall structure of the dual attention module.

perform a matrix multiplication between F_{o1} and the transpose of F_{o2} to obtain the spatial attention weight matrix $W \in \mathbb{R}^{N\times N}$ with the help of the softmax layer:

$$\omega_{j,i} = \frac{\exp(F_{o_1^i} \cdot F_{o_2^j})}{\sum_{i=1}^{N} \exp(F_{o_1^i} \cdot F_{o_2^j})} \tag{7}$$

where $\omega_{j,i} \in \mathbb{R}^{N\times N}$ represents the influence of voxel i on voxel j. In parallel, the input feature $F_{in} \in \mathbb{R}^{C\times S\times H\times W}$ utilizes a $1\times1\times1$ convolution layer to achieve the feature $F_{o3} \in \mathbb{R}^{C\times S\times H\times W}$. This feature is then reshaped to $\mathbb{R}^{N\times C}$. After this, we perform a matrix multiplication between F_{o3} and the transpose of matrix W, making it consistent with the size of F_{in}. The resulting attentional feature map O^s is described in Eq(8).

$$o_j^s = \alpha_s \sum_{i=1}^{N} (\omega_{j,i} \cdot F_{o3}) + F_{in} \tag{8}$$

where α_s is initialized to 0 and learns to assign more weights; N is the same as in Eq(7). The output has a global contextual view and selectively aggregates contextual cues from the target vessels based on the spatial attention map, improving spatial semantic coherence.

Self-attention mechanism concerns mainly:query (q), key (k), value (v), query content (x_q) and key content (x_k), in multi-head self-attention, the output feature y_q can be formulated in Eq(9):

$$y_q = \sum_{m=1}^{M} W_m \left[\sum_{k\in\Omega_q} A_m(q,k,v,x_k) \odot W'_m x_k \right] \tag{9}$$

where M represents the attention head, $A_m\left(q,k,z_q,x_k\right)$ is the weight in the m^{th} attention head, W_m and w'_m are the learnable weights, Ω_q is the reasonable region for computing the output query.

For deformable convolution, the learnable weights are updated based on contetnt of q and relative position. Thus, deformable convolution in attention can be described as Eq(10-11).

$$y_q^{dc} = \sum_{m=1}^{M} W_m \left[\sum_{k\in\Omega_q} A_m^{dc}(q,k,x_q) \odot W'_m x_k \right] \tag{10}$$

$$A_m^{dc}(q,k,x_q) = g\left(k, q + p_m + w_m^{\top} x_q\right) \tag{11}$$

$g()$ here is the same function as in Eq(6).

2) CHANNEL FEATURE FUSION MODULE

The segmentation of blood vessels in CT sequential images has always been difficult due to the overlapping of adjacent tissues and nearby vessels. Convolution kernels are used to extract the image features and are stored in different channels.

In the segmentation model of encoder-decoder, high-resolution information from deep extraction contains global semantic information which helps to find the boundary of thin vessels. Low-stage information is used for image recognition. The multi-scale features of low-stage and high-stage are key to improving the accuracy of segmentation. Therefore, the high-level features with rich semantic information and the low-level features with abundant spatial information are essentially complementary. In Unet and many of its variants, multi-scale features are directly concatenated: $F^{(c)} = concatenate(F_L{}^{(c)}, upsample(F_H^{(c+1)})$ where c represents the current layer, L and H indicate low-level and high-level features. Unfortunately, such a simple operation makes the use of multi-scale information insufficient.

To improve the segmentation ability of the model for microscopic vessels, in this paper we introduce FFM, where the generated 2D low-stage features from down-sampling and high-stage 2D features from up-sampling are concatenated together according to the first dimension. In this way, semantic information is fully utilized and interference in channels has been effectively filtered. This concatenated feature will subsequently be sent to FFM, modeling correlations between

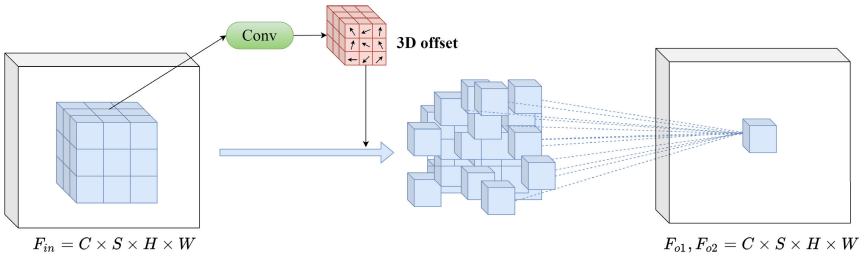


FIGURE 7. Example of 3 × 3 3D deformable convolution.

channels through a squeeze-and-excitation (SE) operation, preserving more semantic context for accurate localization. The new feature can be described in Eq (12):

$$F^{(c)} = F_H^{(c+1)} \oplus (F_L{}^{(c)} \otimes \sigma(GAP(\mathbb{C}_{1\times1}(F_L{}^{(c)})))) \quad (12)$$

where σ represents the sigmoid function $\mathbb{C}_{1\times1}$ represents the 1×1 convolution operation, $\oplus$ and $\otimes$ represents element summing and element multiplication, respectively. We use a parameter-free bilinear up-sampling strategy to reduce the parameters while maintaining segmentation performance and preserving more semantic context for accurate localization. Details can be seen in Fig. 6(c).

C. IterUnet

In this section we apply an iterative network IterUnet, aimed at connecting tiny cracks in CGS. In the following part, we will state the motivation and architecture of this module.

An Iterative strategy has been utilized to get better segmentation results in many existing methods. Even with the same model, iterative application can be of great benefit repairing the gap [39].

The most obvious characteristic of IterUnet is that the input to each iteration is a concatenation of all feature maps generated by all preceding layers, the same structure as dense connection [40]. Suppose that the output of layer $(l-1)^{th}$ is defined as x_{l-1}, the output of l^{th} can be defined as:

$$x_l = \mathbb{U}_l([x_0, x_1, \ldots, x_l - 1]) \quad (13)$$

In IterUnet, we supply penultimate feature from the initial network CSG as input, using Unet as a repeated structure to replace the original dense block. $\mathbb{U}$ here represents a sequence of a series operations in Unet.

One fact we know is that when manually repairing cracks, human infer the connection direction of the broken vessels according to the trend of the vessels in the raw images. Introducing thisapproach into the network will help refine cracks, thus a rated feature is needed to constrain the optimal direction. So we decide to exploit features x_{s1} in skip layers to guide the optimization direction, mainly for the following two reasons. On the one hand, 3D features in the encoder stage cannot be used directly. On the other hand, the low-level x_{s1} realizes the maximum density projection of multiple slice features with the help of a feature dimension transformation, which is closest to the raw images we need. So, Eq(12) can be restated in Eq(14).

$$x_l = \mathbb{U}_l([x_{s1}, x_0, x_1, \ldots, x_{l-1}]) \quad (14)$$

IV. EXPERIMENTS AND RESULTS

In this section, we will give a detailed description of the experimental design, including experiment setup and results analysis.

A. EXPERIMENTAL SETUP

In the case of 1mm thin-cut MIP, we usually have 13-30 slices containing critical vascular information. The purpose of SMA reconstruction is to obtain the projection results of SMA vessels in multiple MIP slices. In this work, we ended up choosing four consecutive MIP slices as inputs to obtain consistent spatial information. Given the morphological differences of SMA across sequential slices, too many slices would result in spatial disturbance. To verify the optimal number of input slices, we set up experiments using 2, 3, 4, 5 slices as input and compared the resulting loss in various models to see how well each input slice sequence converged. The visual loss graph of different input strategies is shown in Fig. 8: the training process gives better performance in model fitting when there are four consecutive slices. Therefore, in our subsequent experiments, input sequences of length 4 are used.

The platform used in this experiment comes from a deep learning computing platform with two NVIDIA RTX-3090 24GB graphics cards. The operating system and version is ubuntu 20.04, the machine learning environment configuration is torch 1.7.0 with CUDA 11.1 and the program compilationing environment is Python 3.6.12.

B. DATASET

CT scans used in the experiment were acquired from a Siemens dual-source CT scanner (Somatom Force, Siemens Healthcare, Forchheim, Germany). We list the scan parameters in Table 1.

From June 2021 to December 2021, we collected 150 patients' Abdominal CT scans from Affiliated Hospital 2 of Nantong University and completed the reconstruction of thin MIP. Scans have a resolution of 512×512 pixels per slice,

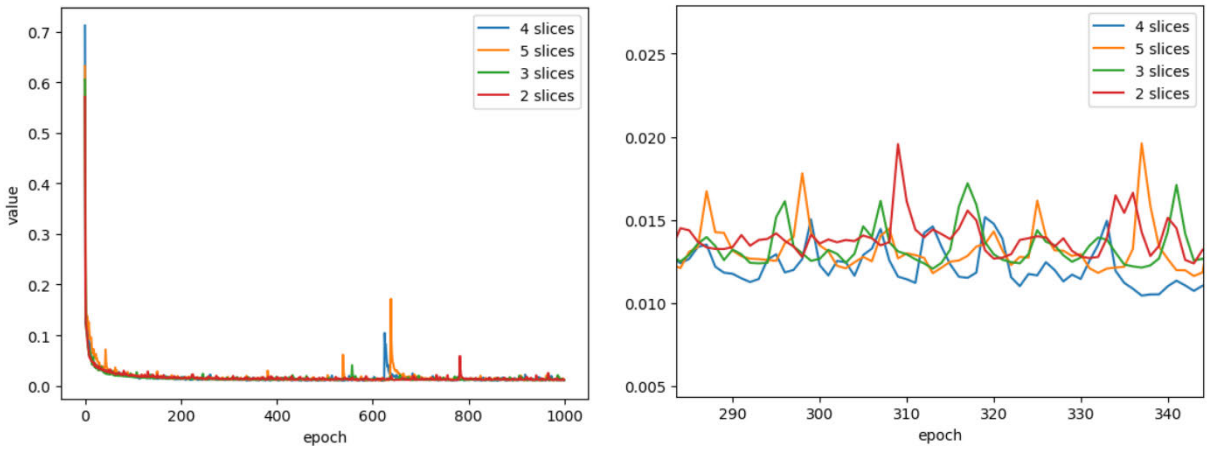


FIGURE 8. Training loss curve (left) and its local enlarged curve (right) for different input strategies in the training process. The 4 adjacent slice input strategy can achieve the least training loss.

TABLE 1. Scan Parameters of the CT Equipment.

Group	Scan parameter
Scanning range	Diaphragmatic apex to inferior margin of symphysis pubis
Position	Supine
Tube voltage	A 90KV/B Sn150KV
Tube current	A 144mAs/B 90mAs
Pitch	1.0
Speed	0.5S
Collimation	2×192×0.6 mm
Contrast medium concentration	370mgI/ml
Slice thickness	1mm
Slice gap	1mm

and each MIP sequence has from 13 to 30 slices. For single group experiment, we have a data set consisting of 2148 MIP slices. Based on the distribution of SMA, we select a series of slices that best represents the characteristics of SMA. For a sequence of length N, N-3 sets of results can be obtained for every 4 consecutive overlapping entries as input. Three experts were invited to manually annotate these vessels. In order to evaluate the effect of the proposed model objectively, we split the MIP dataset into three parts, namely the normal group, the low contrast group, and the large area AA interference group. We employ a 5-fold cross-validation strategy and report the average results, where each fold contains 60% scans for training, 20% for validation, and 20% for testing. To increase the variability of the data, we rotate, flip and mirror the images randomly, but without augmenting the dataset size.

In the subsequent evaluation, we will evaluate the model from two perspectives namely, the reconstruction of a single group and the ensemble group (150 samples). A MIP sequence composed of 15 images can be regarded as one 15-image ensemble group and 12 single (4-image) groups.

C. EXPERIMENT

In this section, we first conduct a comparative experiment on the model backbone, then we compare models with different loss functions to choose a loss function that is more suited to thin MIP reconstruction. Next, we conduct an ablation study of our method to verify the serviceability of the innovative structures. Finally, we compare our work with state-of-the-art approaches on our dataset.

1) BACKBONE STUDIES

We integrate spatial context information into 2D segmentation tasks. To demonstrate the effectiveness of this operation, a comparison between using single image (SIS) and using CGS is conducted. Additionally, simply concatenating low-level features with high-level features introduces interference and noise coming from the background. This fails to fully utilize semantic information. We compare the results of employing channel attention (CGS+ FFM) and direct concatenation (CGS+CAT) from two perspectives: single group and ensemble group, in order to demonstrate the impact of FFM on the above problems.

To quantitatively evaluate the backbone model, we select the F1 score (F1) as the final evaluation index. Meanwhile, scores of Accuracy (A), Precision (P) and recall (R) in the test set are demonstrated.

From Table 2, we observe that CGS+CAT obtains better performance than SIS. By comparing the two different feature extraction structures, the model with CGS obtains higher scores in R, P, and F. The introduction of context information increases the measure of R from 0.63 to 0.684 and F1 from 0.691 to 0.720. In the ensemble group, R and F1 increased to 0.740 and 0.763, 4.3% and 2.1% higher than the original 2D result respectively. Although they have almost the same complexity, their segmentation performance has significant differences. Three-dimensional information integrates higher-dimensional spatial features, which is more robust when reconstructing vascular vessels.

We further analyze the performance of FFM. From Table 2, we observe that CGS+FFM performs better than CGS+CAT. For P, CGS+FFM achieves 0.709 in single group and 0.752 in ensemble stacking group. CGS+CAT achieves a P score of

TABLE 2. Quantitative Results of Different Backbones by Five-Fold Cross-Validation for Datasets (A, P, R, F1). The Best Result Is in Bold Text.

Approach	model	A	P	R	F1
Single	SIS	0.963	0.630	**0.820**	0.691
	CGS+CAT	0.970	0.684	0.788	0.720
	CGS+FFM	**0.972**	**0.711**	0.776	**0.731**
Ensemble	SIS	0.961	0.697	**0.839**	0.742
	CGS+CAT	0.966	0.740	0.812	0.763
	CGS+FFM	**0.968**	**0.752**	0.804	**0.769**

only 0.684 and 0.740 respectively, 2.5% 1.2% lower than CGS+FFM. For F1 score, CGS+FFM achieves 0.735 in single group and 0.769 in ensemble stacking group, which is 1.3% and 0.6% higher than those in CGS+CAT. These results consistently demonstrate that feature fusion using FFM is superior to direct concatenation. P and R are negatively correlated. As a comprehensive metric of P and R, F comes out best in both groups of experiments.

TABLE 3. Mean Results of Model With Different Loss Functions, the Best Result Is Shown in Bold Text.

Approach	Loss	A	P	R	F1
Single	DICE loss	0.964	0.615	**0.834**	0.695
	CE loss	**0.972**	**0.711**	0.776	**0.731**
Ensemble	DICE loss	0.966	0.740	**0.836**	0.754
	CE loss	**0.968**	**0.752**	0.804	**0.769**

2) LOSS FUNCTIONS

We next investigate the impact of the network loss function on reconstruction performance. CE Loss and DICE Loss are respectively used to train models. The quantitative and qualitative results can be found in Table 3 and Fig 9. As can be seen from Table 3, compared with DICE Loss, the model using CE Loss improves by 1% and 4% on A and F1 respectively, while the score of R decreases by 5%, The optimization objective of CE Loss can cause the model to obtain a higher p-measure. For the ensemble group, CE Loss achieves 0.769 in F1 and 0.753 in P, 1.5% and 4.7% higher than those in DICE Loss. These results show that CE Loss can be used to segment more extensive and deeper blood vessels, which are mainly manifested as low-contrast target vessels in the background with similar feature expressions. Recognition of more target vessels increases the P score. Further analysis is conducted by comparing the intuitive segmentation results of these two loss functions, and the final results are displayed in the form of a pseudo-color graph after superposition. Compared with Dice Loss, the gray scale of vessels obtained by CE Loss is wide and the boundary of vessels is fuzzy, so is uncertainty about the pixels of some vessels with low contrast. In general, the brightness of the gray map can be expressed as the probability of being target vessels, so more vascular information is retained (as shown in the arrow in the last column of Fig. 9). By observing columns (a-e) in Fig. 9, we find that CE Loss segments small vessels with low contrast earlier than DICE Loss, and the overall consistency of vessels is stronger. We admit that using CE Loss to identify more target vessels will lead to some mis-segmentation problems. Addressing these questions, we will introduce more sub-modules to deal with these problems in the ablation experiment.

3) ABLATION STUDIES

We present both quantitative and qualitative experiment results of the submodules we proposed above. To avoid disturbance of the results caused by the randomness of data augmentation, we apply a 5-fold-cross-evaluation on our dataset and take the average result of the parameters as the final result. The corresponding data can be seen in Table 4, while the visual segmentation can be found in Fig. 10. We start from baseline (CGS+FFM), then we evaluate the CGS+DA, which employs the Scale-aware attention module for strengthening the extraction of thin vessels while alleviating the distribution of noise and background on the basis of (CGA+FFM). Finally, we process the model with IterUnet (CGS+DA+IterUnet) to optimize vessels with tiny cracks.

We first evaluate the performance of the scale-aware attention block. As depicted in Table 4, the F1 score of CGS+DA is much higher than that of CGS+FFM. For F1, CGS+DA achieves 0.791 in single group and 0.807 in ensemble group, which is 4.8% and 0.6% higher than models without scale-aware attention blocks. The high score of a single group indicates stability of the model. The result of the ensemble group is obtained by superposition of single groups. If a target vessel is not separated in a single group, but is separated in a subsequent single group, it has almost no impact on the result of the ensemble group. In the results for multiple single groups of a patient, there are many overlapping areas of main vessels, so the tiny marginal vessels with a small proportion in a single group determine the vascular extension range and vascular accuracy of the ensemble group. Therefore, although the improvement of the ensemble group performance is not statistically outstanding, it can reflect better reconstruction performance.

In particular, we noticed the increase in P score. For P, the addition of the Scale-aware attention module increased by 7.9% and 1.4% in the single group and the ensemble respectively. The improvement of P here is fairly critical, and the visual effect is demonstrated in Fig. 10 (Ensemble group). We first notice that the proposed DA enhanced the segmentation ability of thin vessels in the end. This effect

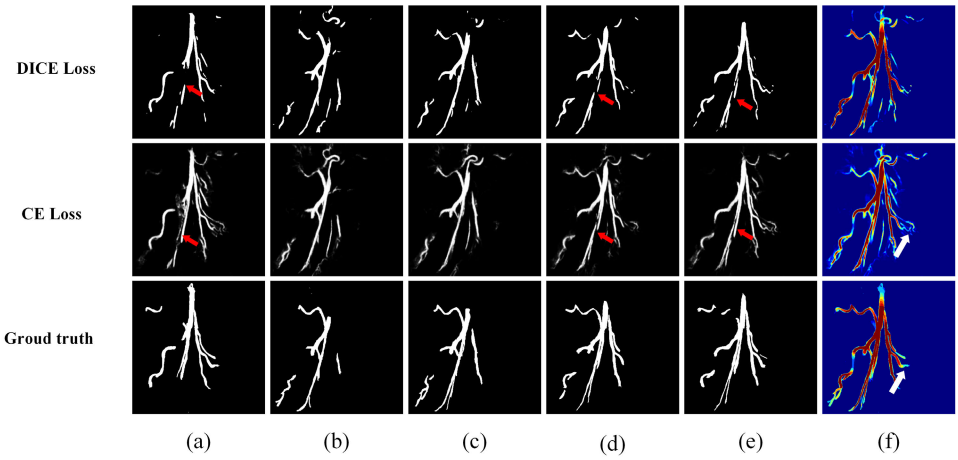


FIGURE 9. Performance comparison between loss functions. The five columns from left to right show the results of each of the four inputs, and the last column (f) displays the stacking result of the whole sequence. Columns from top to bottom indicate model with DICE Loss, model with CE Loss, and ground truth annotation.

TABLE 4. Quantitative Results for Ablation Studies. All the Methods Are Based on CGS.

Approach	model	A	P	R	F1
Single	CGS+FFM	0.972	0.709	0.785	0.735
	CGS+DA	0.978	0.788	0.806	0.791
	CGS+DA+IterUnet	**0.978**	**0.811**	**0.792**	**0.798**
Ensemble	CGS+FFM	0.968	0.752	0.804	0.769
	CGS+DA	0.976	0.798	0.822	0.807
	CGS+DA+IterUnet	**0.976**	**0.799**	**0.824**	**0.808**

benefits from the adaptively adjusted receptive field in the module, which can adaptively focus on small blood vessels. Meanwhile, we note that adjacent vessels with similar feature expressions and artifacts in adjacent vessels have been greatly improved. This effect benefits from the multi-use of channel attention mechanism, which causes attention features to focus on the most discriminating regions.

We finally investigate the effectiveness of the proposed IterUnet. In Table 4, we observe that the addition of IterUnet doesn't make a big change to scores as it only works on tiny vessels. The tiny change can be seen in the last column in Fig. 10, where we apply false color to render local enlargement performance: the darker the color, the better the continuity of the blood vessels. By introducing weight-sharing in IterUnet, we successfully empower the model with the ability to find possible defects in the intermediate results and fix them in a reasonable way. The experimental results prove that the proposed module has been successful.

4) COMPARISON WITH OTHER STATE-OF-THE-ART METHODS

We compared our models with some state-of-the-art ones including original Unet, S-UNet [41], R-net [42] and AA-UNet [10]. The latter three are proposed for retinal segmentation. Visual results can be seen in Fig. 11 (single group).

We notice that all of these methods can extract the rough outline of vessel trees. The main difference lies in the segmentation performance of the terminal small vessels and the processing of background noise and interference. Comparing our method with S-UNet, (see the comparison between the third and fourth columns), from a series of predicted datasets, we select two images that best represent the segmentation problem commonly seen in S-UNet. From the local magnified image, it is found that the problems with the treatment of edge vessels are as follows (see Fig. 11): blurred vessels, poor continuity, loss of small foreground vessels (red arrow), and improper segmentation of background vessels under low contrast (blue arrow). Compared with thick vessels, thin vessels with low contrast are extremely similar to artifacts in the background, which increases the challenge for accurate segmentation. When manually labeled, it is difficult to distinguish the attributes of fine vessels with the naked eye. One tip for doctors to determine whether they belong to the foreground or background is to track the main vessels of fine vessels.

Then we analyze the performance of R-Net. On the whole, the segmentation result of R-Net is much better than that of S-UNet which mainly reflected in the clear vessel tree and stable continuity of vascular structures. Similar to the problem with S-UNet, R-Net fails to segment deeper blood vessels (green arrow). In addition, R-Net brings about partial background disturbance since it lacks global guided contextual information, which is manifested by the appearance of other vessels dissociating from the vessel tree(blue arrow). This phenomenon is depicted in the unamplified segmentation of each sample. The remaining comparison networks,

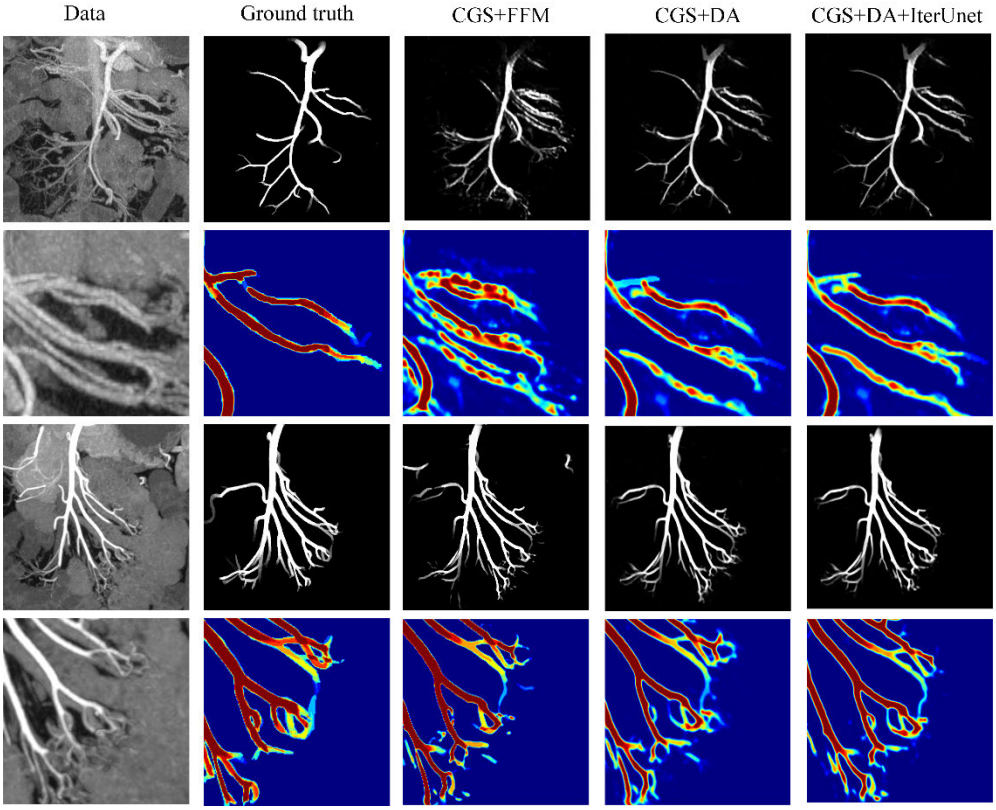


FIGURE 10. Qualitative comparison of the baseline model, dual attention guided module, and IterUnet module. From left to right, each column represents the one slice in sequential input, ground truth, segmentation results of CGS+FFM, segmentation results with the guidance of dual-attention module and segmentation output with IterUnet.

TABLE 5. Quantitative Results of the State-of-the Art Models by Cross-Validation on Our Dataset With Respect to F1 Score.

		F1 score		
Approach	model	normal	Low contrast	AA Interference
Single	Unet	0.725	0.659	0.679
	S-UNet	0.760	0.686	0.693
	X-ray_net	0.792	0.754	0.736
	R-Net	0.748	0.733	0.718
	AA-UNet	0.802	0.754	0.732
	Our Model	**0.823**	**0.768**	**0.747**
Ensemble	Unet	0.753	0.687	0.712
	S-UNet	0.804	0.675	0.725
	X-ray_net	0.825	0.701	0.753
	R-Net	0.746	0.686	0.718
	AA-UNet	0.829	0.763	0.756
	Our Model	**0.857**	**0.786**	**0.794**

Unet, X-ray net and AA-UNet can merely segment the main structure of SMA, and struggle to distinguish the tiny vessels which are only a few pixels wide. Finally, we compare the result of the raw images with similar interference, by comparing the pictures in the last row of Fig. 11. It can be seen that besides our model, only X-ray has successfully learned and utilized context information to remove the interference. Unfortunately, the accuracy of the X-ray net is inadequate. From the quantitative perspective, our model comes top in all three datasets which can be found in Table 5. All other models just realize a mapping of the original images.

Experiments show that our proposed model segments almost the complete range of vessels, effectively suppressing the disturbance of low contrast and artifacts, and ensuring the continuity of vessels. It reduces the problem of broken blood vessels, and obtains the best overall segmentation results.

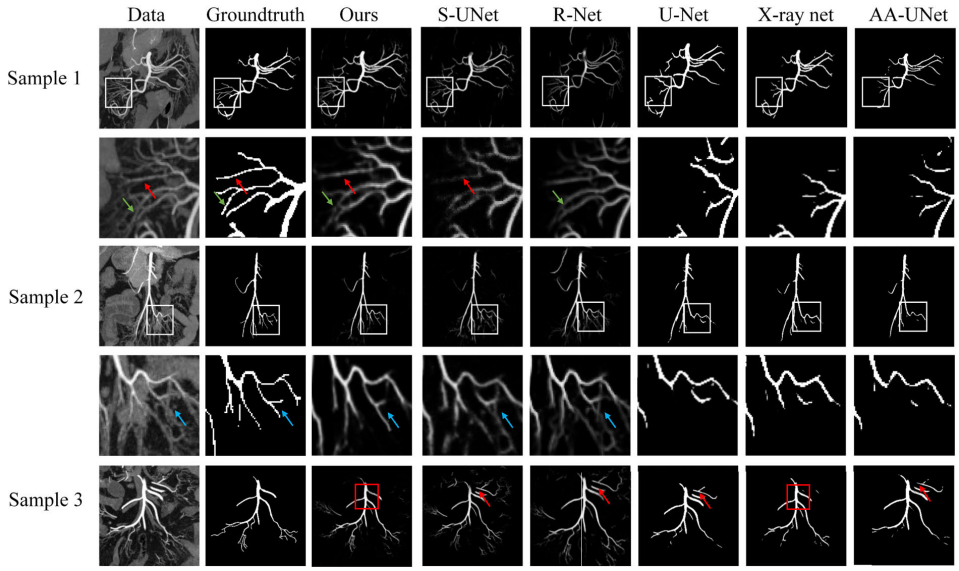


FIGURE 11. Qualitative segmentation results by different vessel segmentation methods. From left to right, each row displays:the raw image, the manually outlined ground truth (even rows are a magnified view of the local region of odd rows), the result of methods segmented by our method, SU-Net, R-Net, Unet, X-ray and AA-UNet net respectively.

V. DISCUSSION AND CONCLUSION

Currently, most existing indicators for vessel segmentation are pixel-based measurement, where thin vessels may not contribute much to the value of the indicators as they contain only a small number of pixels. However, in the field of medical imaging, the pixel-based metrics mentioned above may experience the problem of class imbalance. In this work, we added adaptive attention mechanism to make the model more actively focus on the characteristics of the blood vessels themselves, especially small blood vessels. However, this performance improvement comes at the cost of a larger parameter space, how to lightweight the network will be explored as part of our future work.

In this work, we present a novel architecture for the reconstruction of SMA based on MIP sequence. In particular, we propose a novel backbone CGS. In this structure, to include contextual information we take a sequence of consecutive images in the MIP sequence as the input, and use the superimposed image of the target vessel in these images as the output. The mixed use of 3D and 2D convolution introduces spatial-temporal features into the segmentation task, succeeding in achieving relatively complete vessels which can be displayed on screen rather than relying on an artificial combination. Two sub-modules, a multi-scale aware attention module and FFM, are embedded in our CGS. The former can adaptively extract features from multi-scale feature maps. This operation helps us to progressively aggregate relevant contextual features and guides us to focus on smaller features. The latter helps filter disturbance from different channels, reducing the interference from peripheral vessels.

To validate the best input sequence length, we conduct experiments on different input lengths and finally confirm 4 slices as the optimal context. Then we further prove the structural advantages and the rationality of using FFM to fuse features. We next provide ablation experiments to evaluate the impact of the individual components of the proposed architecture. Last, we compare the model with approaches that have been recently proposed. Experimental results show that the proposed model outperforms all previous approaches both quantitatively and qualitatively, largely due to the adaptive ability to model rich contextual dependencies. This demonstrates the efficiency of our approach in providing precise and reliable automatic segmentation of MIP sequences. We believe that this approach is flexible and can be extended to other MIP sequence tasks where complete coronal imaging is required.

REFERENCES

[1] Z. Zhang, X. Chen, C. Li, H. Feng, H. Yu, and R. Zhu, "Percutaneous mechanical thrombectomy for acute superior mesenteric artery embolism: Preliminary experience in five cases," *Ann. Vascular Surg.*, vol. 63, pp. 186–192, Feb. 2020.

[2] J. Wang, H. Zhu, S.-H. Wang, and Y.-D. Zhang, "A review of deep learning on medical image analysis," *Mobile Netw. Appl.*, vol. 26, no. 1, pp. 351–380, 2021.

[3] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2017.

[4] A. Fourcade and R. H. Khonsari, "Deep learning in medical image analysis: A third eye for doctors," *J. Stomatol., Oral Maxillofacial Surg.*, vol. 120, no. 4, pp. 279–288, Sep. 2019.

[5] S. Zheng, X. Cui, M. Vonder, R. N. J. Veldhuis, Z. Ye, R. Vliegenthart, M. Oudkerk, and P. M. A. van Ooijen, "Deep learning-based pulmonary nodule detection: Effect of slab thickness in maximum intensity projections at the nodule candidate detection stage," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105620.

[6] R. Cao, Y. Jiang, J. Lu, G. Wu, L. Zhang, and J. Chen, "Evaluation of intracranial vascular status in patients with acute ischemic stroke by time maximum intensity projection CT angiography: A preliminary study," *Academic Radiol.*, vol. 27, no. 5, pp. 696–703, May 2020.

[7] L. Pei, W.-W. Hsu, L.-A. Chiang, J.-M. Guo, K. M. Iftekharuddin, and R. Colen, "A hybrid convolutional neural network based-method for brain tumor classification using mMRI and WSI," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2020, pp. 487–496.

[8] H. Pinckaers, W. Bulten, J. van der Laak, and G. Litjens, "Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels," *IEEE Trans. Med. Imag.*, vol. 40, no. 7, pp. 1817–1826, Jul. 2021.

[9] S. Trajanovski, C. Shan, P. J. C. Weijtmans, S. G. B. de Koning, and T. J. M. Ruers, "Tongue tumor detection in hyperspectral images using deep learning semantic segmentation," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 4, pp. 1330–1340, Apr. 2021.

[10] Y. Lv, H. Ma, J. Li, and S. Liu, "Attention guided U-Net with atrous convolution for accurate retinal vessels segmentation," *IEEE Access*, vol. 8, pp. 32826–32839, 2020.

[11] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[12] B. Wang, S. Qiu, and H. He, "Dual encoding U-Net for retinal vessel segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 84–92.

[13] L. N. L. Thuy, T. D. Trinh, L. H. Anh, J. Y. Kim, H. T. Hieu, and P. T. Bao, "Coronary vessel segmentation by coarse-to-fine strategy using U-nets," *BioMed Res. Int.*, vol. 2021, pp. 1–10, Apr. 2021.

[14] C. Ma, G. Luo, and K. Wang, "Concatenated and connected random forests with multiscale patch driven active contour model for automated brain tumor segmentation of MR images," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1943–1954, Aug. 2018.

[15] M. J. Cardoso, C. H. Sudre, M. Modat, and S. Ourselin, "Template-based multimodal joint generative model of brain data," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2015, pp. 17–29.

[16] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[17] J. Sun, Y. Peng, Y. Guo, and D. Li, "Segmentation of the multimodal brain tumor image used the multi-pathway architecture method based on 3D FCN," *Neurocomputing*, vol. 423, pp. 34–45, Jan. 2021.

[18] E. Sanderson and B. J. Matuszewski, "FCN-transformer feature fusion for polyp segmentation," in *Proc. Annu. Conf. Med. Image Understand. Anal.* Cham, Switzerland: Springer, 2022, pp. 892–907.

[19] H. Zhang, X. Tang, and X. Han, "High-resolution remote sensing images change detection with Siamese holistically-guided FCN," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 4340–4343.

[20] S. Kim, S. An, P. Chikontwe, and S. H. Park, "Bidirectional RNN-based few shot learning for 3D medical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 1808–1816.

[21] Y. Qi and Z.-H. Tan, "SketchSegNet+: An end-to-end learning of RNN for multi-class sketch semantic segmentation," *IEEE Access*, vol. 7, pp. 102717–102726, 2019.

[22] L. H. Shehab, O. M. Fahmy, S. M. Gasser, and M. S. El-Mahallawy, "An efficient brain tumor image segmentation based on deep residual networks (ResNets)," *J. King Saud Univ.-Eng. Sci.*, vol. 33, no. 6, pp. 404–412, Sep. 2021.

[23] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[24] F. Zmilletari, N. Rieke, M. Baust, M. Esposito, and N. Navab, "CFCM: Segmentation via coarse to fine context memory," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 667–674.

[25] Y. Jiang, S. Xu, H. Fan, J. Qian, W. Luo, S. Zhen, Y. Tao, J. Sun, and H. Lin, "ALA-Net: Adaptive lesion-aware attention network for 3D colorectal tumor segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3627–3640, Dec. 2021.

[26] A. D. Huberman, M. B. Feller, and B. Chapman, "Mechanisms underlying development of visual maps and receptive fields," *Annu. Rev. Neurosci.*, vol. 31, p. 479, 2008.

[27] C. Li, D. Zhang, Z. Tian, S. Du, and Y. Qu, "Few-shot learning with deformable convolution for multiscale lesion detection in mammography," *Med. Phys.*, vol. 47, no. 7, pp. 2970–2985, Jul. 2020.

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

[29] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.

[30] X. Liang, N. Li, Z. Zhang, J. Xiong, S. Zhou, and Y. Xie, "Incorporating the hybrid deformable model for improving the performance of abdominal CT segmentation via multi-scale feature fusion network," *Med. Image Anal.*, vol. 73, Jan. 2021, Art. no. 102156.

[31] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-Net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020.

[32] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10076–10085.

[33] A. Moyes, R. Gault, K. Zhang, J. Ming, D. Crookes, and J. Wang, "Multi-channel auto-encoders for learning domain invariant representations enabling superior classification of histopathology images," *Med. Image Anal.*, vol. 83, Jan. 2023, Art. no. 102640.

[34] M. Pominova, E. Kondrateva, M. Sharaev, A. Bernstein, S. Pavlov, and E. Burnaev, "3D deformable convolutions for MRI classification," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 1710–1716.

[35] J. Li, L. Huang, Z. Wei, W. Zhang, and Q. Qin, "Multi-task learning with deformable convolution," *J. Vis. Commun. Image Represent.*, vol. 77, May 2021, Art. no. 103109.

[36] Z. Shi, X. Liu, K. Shi, L. Dai, and J. Chen, "Video frame interpolation via generalized deformable convolution," *IEEE Trans. Multimedia*, vol. 24, pp. 426–439, 2022.

[37] Y. Haiying, F. Zhongwei, D. Ding, and S. Zengyang, "False-positive reduction of pulmonary nodule detection based on deformable convolutional neural networks," in *Proc. IEEE 9th Int. Conf. Bioinf. Comput. Biol. (ICBCB)*, May 2021, pp. 130–134.

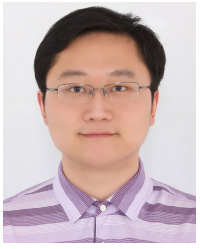
[38] T. Dou, X. Yu, and J. Zhao, "Multi-kernel deformable 3D convolution for video super-resolution," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Aug. 2021, pp. 1–6.

[39] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki, "IterNet: Retinal image segmentation utilizing structural redundancy in vessel networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3656–3665.

[40] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[41] J. Hu, H. Wang, S. Gao, M. Bao, T. Liu, Y. Wang, and J. Zhang, "S-UNet: A bridge-style U-Net framework with a saliency mechanism for retinal vessel segmentation," *IEEE Access*, vol. 7, pp. 174167–174177, 2019.

[42] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369–2380, Mar. 2016.



KUN ZHANG (Member, IEEE) was appointed as the Vice Chair with the Nantong Key Laboratory of Intelligent Control and Intelligent Computing, in 2021. He has been a Key Member with the Nantong Key Laboratory of Intelligent Medicine Innovation and Transformation, since 2021. He is currently a Professor with the School of Electrical Engineering, Nantong University, Nantong, China. He has published over 60 scientific papers in journals and international conferences. His current research interest includes machine learning for computer vision.

YU HAN is currently pursuing the master's degree in control science and engineering with the School of Electrical Engineering, Nantong University. His current research interests include pattern recognition and medical images.

PEIXIA XU received the B.S. degree in electrical engineering from Nantong University, Nantong, China, in 2021, where she is currently pursuing the M.S. degree. Her research interests include medical imaging, computer vision, and other related research fields.

MEIRONG WANG received the master's degree in imaging medicine and nuclear medicine from Nantong University, in 2018. She is currently working with the Imaging Department, Second Affiliated Hospital, Nantong University. She presided over the youth project of Nantong Health Committee and participated in the implementation of many projects. Her current research interests include imaging diagnosis of abdominal diseases and machine learning of mesenteric vessels.

JUSHUN YANG is currently the Deputy Chief Physician with the Imaging Department, Second Affiliated Hospital of Nantong University. He has been engaged in imaging diagnosis (X-ray, CT, and MRI) for more than ten years. He has published more than ten scientific papers in journals and international conferences. His research interest includes abdominal imaging diagnosis.

PENGCHENG LIN is currently pursuing the master's degree in control engineering with the School of Electrical Engineering, Nantong University, China. His research interests include medical image division, deep learning, and machine vision.

DANNY CROOKES (Senior Member, IEEE) was appointed as the Chair of computer engineering with Queens University Belfast, in 1993, and the Head of computer science, from 1993 to 2002. He is currently an Emeritus Professor of computer science with Queens University Belfast, U.K. He has published over 260 scientific papers in journals and international conferences. His current research interests include machine learning for computer vision and the use of novel architectures (especially FPGAs) for high-performance image processing.

BOSHENG HE received the M.D./Ph.D. degree in imaging medicine and nuclear medicine from Soochow University, in 2015. He is currently the Deputy Chief Physician with the Imaging Department, Second Affiliated Hospital of Nantong University, and the Director of the Nantong Key Laboratory of Intelligent Medicine Innovation and Transformation. His research interests include imaging diagnosis of abdominal related diseases, intestinal ischemia, and vascular artificial intelligence.

LIANG HUA was appointed as the Head of the School of Electrical Engineering, from 2017 to 2021. He is currently a Professor of computer science and robot intelligence with Nantong University, Nantong, China. He has published over 40 scientific papers in journals and international conferences. His current research interests include machine vision and smart robot control.