TPCM-SegNet: A Text-Prompted Dual-Path Convolution-Mamba Network for Anomaly Segmentation

Borong Xu, Junyu Dong, *Member, IEEE* and Xinghui Dong, *Member, IEEE*

***Abstract*—Anomaly segmentation has been widely applied to diagnosis of medical organs and lesions and detection of industrial defects. However, existing methods still face challenges in extracting discriminant image features and utilizing semantic information. To address these issues, we propose a Text-Prompted Dual-Path Convolution-Mamba Network (TPCM-SegNet)[1], which integrates Residual Double-Convolution Blocks (RDCBs) and Mamba-Transformer Blocks (MTBs) in two parallel paths for the purpose of extracting local and global features, respectively. Given a pair of RDCB and MTB at the same stage, a Feature Fusion Block (FFB) is introduced in order to facilitate the interaction and fusion of the features extracted using these blocks. Furthermore, we fuse the text tokens extracted from a textual description with the image features extracted using each of those blocks through a Text Prompt Block (TPB), to enhance the semantics understanding ability of the network. A Cascade Feature Block (CFB) is also designed for each stage of the encoder, to combine the feature maps, the logit maps decoded from them and the input image. This block incorporates the prior and original characteristics into the image representation. Experimental results demonstrate that our TPCM-SegNet achieves the superior, or at least comparable, performance to baselines, across eight publicly available datasets. These promising results should benefit from the powerful ability of image representation and semantic understanding of the proposed network.**

***Index Terms*—Anomaly Segmentation, Image Segmentation, Defect Detection, Text Prompt, Mamba**

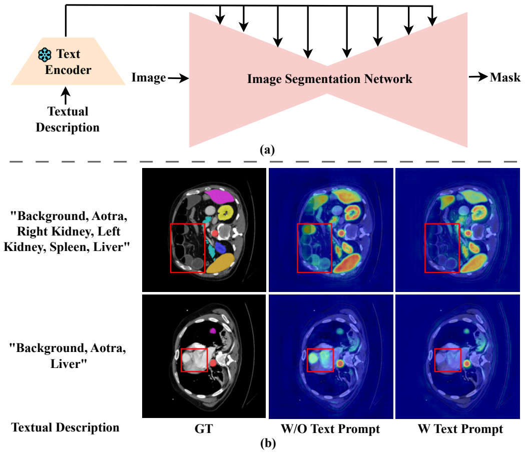
Text Encoder
Textual Description
Image
Image Segmentation Network
Mask
(a)
"Background, Aotra, Right Kidney, Left Kidney, Spleen, Liver"
"Background, Aotra, Liver"
Textual Description
GT
W/O Text Prompt
W Text Prompt
(b)

Fig. 1. Illustration of the importance of text prompt to image segmentation. In (a), we present the preliminary structure of a text-prompted image segmentation network. In (b), the results produced by our segmentation network without and with a text prompt are compared. It can be seen that the text-prompted segmentation network can more accurately focus and identify the object that the text describes than the network which does not use this information.

I. Introduction

ANOMALY segmentation [1], [2] aims to detect anomalous objects, such as defects and lesions, in images. It has been widely applied in medical image processing for detection of organs and lesions [3], [4], [5], and in industrial applications for identifying defects [6], [7]. However, challenges remain in anomaly segmentation, such as extraction of powerful feature representation and utilization of the semantic information of images.

Due to the limited receptive field size of Convolutional Neural Networks (CNNs), existing methods [8], [9], [10] were normally focused on extracting local features. As a result, they struggled with capturing long-range dependencies. In contrast, Transformer [11] was able to utilize these information and has been applied to anomaly detection tasks [12], [13], [14], [15], [16], [17]. However, the quadratic complexity of the self-attention mechanism that Transforms use in terms of the sequence length poses challenges for training and deploying them.

Recently, the Mamba [18] structure, which was designed based on the State Space Model (SSM), emerged as a sequence modeling approach and has been gradually introduced into vision tasks [19], [20]. Since Mamba [18] performs implicit global modeling through recursive computation, they can effectively capture long-range dependencies while significantly reducing computational complexity.

To better exploit local and global characteristics, the joint use of CNN and Transformer [11] or Mamba [18] has attracted attention from researchers. For example, Chen et al. [12] adopted a hybrid encoding approach on top of CNN and Transformer. To capture the broader contextual information and multi-scale characteristics, Xu [21] integrated Mamba with dilated convolution and depth-wise separable convolution. Inspired by MambaVision [22], Zhang et al. [5] introduced a

This study was supported in part by the National Natural Science Foundation of China (NSFC) (No. 42176196) and in part by the Key Research and Development Program of Shandong Province, China (No. 2024ZLGX06) (Corresponding author: Xinghui Dong).

B. Xu, J. Dong and X. Dong are with the State Key Laboratory of Physical Oceanography and the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, 266100. (e-mail: xuborong@stu.ouc.edu.cn, dongjunyu@ouc.edu.cn, xinghui.dong@ouc.edu.cn).

[1] The code and models will be available at https://indtlab.github.io/projects/TPCM-SegNet upon the acceptance of the paper.

MambaVision-based UNet [8], which had a hybrid Mamba-Transformer and CNN structure. Nevertheless, these methods normally extract local and global features sequentially, which lack effective interaction and fusion.

On the other hand, the semantics of images represented in the form of text was rarely used by existing anomaly segmentation methods even if this information had been utilized in other vision tasks [23], [24], as depicted in Fig. 1. In [25], semantic maps were extracted from feature maps using a semantic attention mechanism. But they only contained the relatively weak semantic information and were not intuitive. Recently, the gap between images and text has been bridged due to the introduction of Contrastive Language–Image Pre-training (CLIP) [26]. Although text tokens can be extracted using a text encoder and can be used as a prompt for a vision task instead of directly using the pre-trained large image-text pair model [27], [28], the alignment operation between image features and them is still challenging. The challenge mainly lies in the inconsistency of semantic granularity and feature space between the two modalities. Specifically, text tokens are often abstract and contain high-level semantics, while image features usually focus on more spatial details and local structures. As a result, simple feature fusion methods become ineffective for achieving precise image-text alignment.

In addition, features are extracted layer by layer during the encoding phase, which inevitably leads to the loss of detailed information. Residual connections were proposed for ResNet [29], to inject low-level features into higher-level features. In the depth estimation task, Zhang et al. [30] used a geometric fusion network to extract geometric priors from coarse estimations, which improved the context representation of the following stages. Within the field of image segmentation, challenges remain in effectively leveraging prior characteristics, even though some studies [12], [31], [32] have been focused on improving feature representation using contextual information. In particular, shallow features typically contain rich edges, textures and other details in the encoding stage. Effectively using these features to obtain prior characteristics and integrating them with contextual information can further improve the representation of deep features.

To address the aforementioned issues, we propose a novel Text-Prompted Dual-Path Convolution-Mamba Network for anomaly segmentation, i.e., TPCM-SegNet. This network is built on top of an encoder-decoder architecture with skip connections. One path comprises a series of Residual Double-Convolution Blocks (RDCBs), which aim to extract local features. The other path consists of a set of Mamba-Transformer Blocks (MTBs), which are used to extract global features. Each MTB corresponds to an RDCB at the same stage. For the purpose of injecting the local or global features into the other side, we design a Feature Fusion Block (FFB), which allows the interaction between the two paths. To enable the network to understand the semantics of images encoded in the form of text, we further introduce a Text Prompt Block (TPB), which fuses the text tokens extracted from a textual description with the image features extracted at each RDCB or MTB. Besides, we adopt a Cascade Feature Block (CFB) for each stage of the encoder. This block concatenates the feature maps, the logit maps decoded from these maps and the input image. In essence, the CFB brings prior characteristics and original image characteristics together with the image representation.

To our knowledge, none of existing image segmentation studies have explored such a dual-path network. Our contributions can be summarized as threefold.

- We introduce a Text-Prompted Dual-Path Convolution-Mamba Network for anomaly segmentation, referred to as TPCM-SegNet. The two paths are built on top of Residual Double-Convolution Blocks (RDCBs) and Mamba-Transformer Blocks (MTBs) that we design, which aim to extract local and global features, respectively. To facilitate the fusion of the two types of features, a Feature Fusion Block (FFB) is proposed, which enables the interaction between both paths and achieves the more discriminant image representation.
- We propose a Text Prompt Block (TPB), which fuses text tokens with the image features extracted at different stages. This block improves the semantics understanding capability of the network and boosts its performance.
- We design a Cascade Feature Block (CFB), which combines prior characteristics and original image characteristics with image representation. As a result, the image representation becomes more powerful.

The remainder of this paper is organized as follows. We review the related work in Section II. In Section III, the proposed methodology is introduced. Experimental setup and results are presented in Sections IV and V, respectively. Finally, we draw our conclusion in Section VI.

II. Relate Work

A. CNN-Based Segmentation Methods

Convolutional Neural Networks (CNNs) have made significant breakthroughs in the field of anomaly segmentation. Since Fully Convolutional Networks (FCNs) [33] were introduced, the encoder-decoder architecture has dominated the field of image segmentation. In particular, UNet [8] and its variants [12], [13], [6] received much attention because they were able to enhance the spatial resolution and optimize boundary details. For example, ERDUNet [34] was developed based on the collaboration of two separate networks, to extract local features and global continuity information. Zou et al. [10] integrated the multi-scale features extracted from the encoder-decoder network in order to fulfill crack segmentation. Due to the limited size of receptive fields that CNNs utilize, however, they struggle to capture long-range dependencies.

B. Transformer-Based Segmentation Methods

Motivated by the success of Transformer [11] in Natural Language Processing (NLP) due to its multi-head self-attention mechanism, which can capture long-range dependencies, Vision Transformer (ViT) [35] has been widely applied to computer vision. Huang et al. [36] adopted the Enhanced Transformer Block and used the Enhanced Transformer Context Bridge to capture both long-range dependencies and local characteristics. Rahman and Marculescu [17] adopted a cascaded and parallel Transformer architecture, which enabled strong multi-scale generalization. Qiu et al. [16] introduced a spatially dynamic multi-head attention mechanism and a deformable patch embedding strategy, which effectively handled heterogeneous target appearances. In [12], TransUNet was proposed by integrating Transformer with the UNet [8] architecture. Considering that the self-attention mechanism incurred significant computational overhead because of its high complexity, Liu et al. [37] introduced Swin-Transformer by designing a sliding-window mechanism. Inspired by the high efficiency of Swin-Transformer, Swin-UNet [13] was adopted by bringing together this technique with UNet [8].

Although the computational complexity issue can be alleviated using the sliding-window mechanism, the networks built on top of Swin-Transformer cannot capture local characteristics. Qi et al. [6] proposed a dual-path U-shaped [8] network, referred to as ICCT-UNet, which contained a CNN path and a window-based Transformer path. The ICCT-UNet captures both local characteristics and long-range dependencies. Nevertheless, a simple CNN subnetwork was used to fuse the results of the two paths. In contrast, our method adopts a direct fusion approach, which further promotes feature fusion. Moreover, our method combines prior characteristics and original image characteristics during the encoding stages, enhancing the representation ability of the network.

C. Mamba-Based Segmentation Methods

Recently, Mamba [18] was proposed on top of the State Space Model (SSM) for the purpose of capturing the global context. Compared with Transformer, Mamba achieves linear computational complexity. Furthermore, VMamba [19] and Vision Mamba [20] extended the application of Mamba to the field of computer vision. For instance, VM-UNet [4] was developed by combining VMamba [19] with UNet [8]. Wang et al. [38] proposed a hierarchical and bidirectional Mamba module, which improved the ability of Mamba to capture both global and local characteristics.

To derive the more powerful feature representation, many studies brought Mamba together with CNN or Transformer. Xu [21] integrated Mamba with dilated convolution and depthwise separable convolution, which enabled high-performance image processing while maintaining the lower computational cost. In [22], a MambaVision block was proposed, which improved the accuracy throughput of the network, compared to the original Mamba architecture. Furthermore, HMT-UNet [5] was introduced based on the MambaVision [22] blocks, CNN blocks and Transformer blocks.

Despite the combination of Mamba, CNN and Transformer might produce the better results, the mutual feature injection between them was often ignored. In contrast, the dual-path network that we propose comprises a CNN path and a Mamba path, which extract local and global features, respectively. A Feature Fusion Block (FFB) is also designed to enable the interaction and fusion between the features extracted using the two paths.

D. Semantics- and Text-Prompted Segmentation Methods

The semantic information of an image refers to the data related to objects, scenes, background and other meaningful elements contained in the image. Regarding image segmentation, these information can help the segmentation method determine the category to which each pixel belongs. To improve the performance of image segmentation, existing methods focused on capturing the semantic-level contextual information and integrating this information with the image-level context. In [39], ISNet captured the contextual information from the category region through a Semantic-Level Context Module (SLCM). Jain et al. [25] extracted a semantic map using the semantic attention mechanism during the encoding stage.

On the other hand, the semantic information encoded in the form of textual descriptions is more intuitive. Therefore, an increasing number of studies leveraged image-text pairs for cross-modal learning, which usually produced the stronger feature representation. Xu et al. [40] and Wu et al. [41] used contrastive loss to ensure the alignment between the image and the text. With the introduction of CLIP [26], the boundary between images and text was broken. As a result, the two modals can be better matched. Many open-vocabulary image segmentation methods [27], [28] were developed on top of the image and text encoders of CLIP [26]. However, these methods mainly relied on the weights of CLIP [26] pre-trained on natural images and the entanglement between its image and text encoders, resulting in the suboptimal performance in domain-specific tasks. In contrast, we design a Text Prompt Block (TPB) in order to inject the semantic information extracted from a textual description into the image features. Since this block can be trained along with the network, it is able to better perform the image-text alignment task and thus inclines to boost the performance of the network.

III. Text-Prompted Dual-Path Convolution-Mamba Network

Considering that existing approaches normally encounter challenges in extracting discriminant image features and utilizing semantic information, we propose a Text-Prompted Dual-Path Convolution-Mamba Network to address these issues. The architecture of the proposed network is shown in Fig. 2(a). As can be seen, the network comprises two parallel paths, which are used to extract local and global features. The two paths are built on top of Residual Double-Convolution Blocks (RDCBs) and Mamba-Transformer Blocks (MTBs), respectively, along with skip connections. We also design a Feature Fusion Block (FFB) (see Fig. 2(d)) to facilitate the interaction and fusion of the features extracted using the RDCB and MTB at the same stage. To strengthen the semantics understanding ability of the network, we further introduce a Text Prompt Block (TPB) (see Fig. 2(b)), which fuses the features extracted from a textual description with the image features extracted using the RDCB or MTB. In addition, a Cascade Feature Block (CFB) (see Fig. 2(c)) is designed for each stage of the encoder. This block combines the feature maps, the logit maps decoded from these maps and the input image. In essence, the CFB injects prior characteristics and original image characteristics into the image representation.

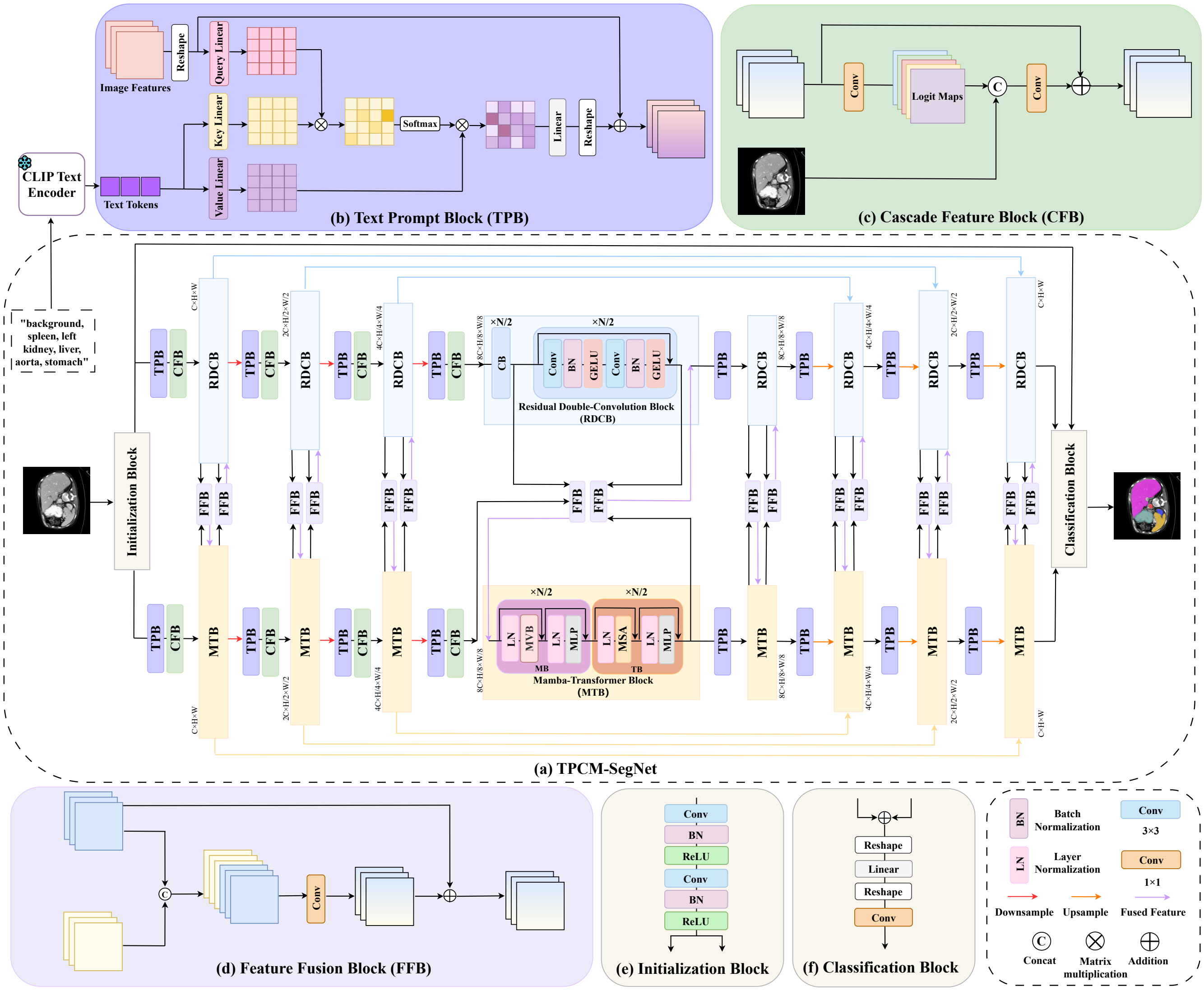

Fig. 2. The architecture of the Text-Prompted Dual-Path Convolution-Mamba Network (TPCM-SegNet) for anomaly segmentation. Here, (a) presents the overall structure of TPCM-SegNet, while (b), (c) and (d) illustrate the detailed structures of TPB, CFB and FFB, respectively. In addition, (e) and (f) show the structures of the initialization block and the classification block, respectively.

A. *Overview*

As shown in Fig. 2(a), the proposed TPCM-SegNet contains a symmetric dual-path encoder-decoder, in which the CNN path comprises TPBs, CFBs and RDCBs while the Mamba path consists of TPBs, CFBs and MTBs. The TPCM-SegNet consists of four symmetric stages in the encoder and decoder, separately. Given a stage, the RDCB or MTB contains N components. In terms of the four stages, the value of N is set to 2, 4, 4, 8 in turn. A skip connection is applied between the two RDCBs or MTBs at the symmetric stages.

Given an image $I \in \mathbb{R}^{3\times H\times W}$, it is first converted into initial features $X \in \mathbb{R}^{C\times H\times W}$ and $Y \in \mathbb{R}^{C\times H\times W}$, using the initialization block (see Fig. 2(e)), which consists of two consecutive 3×3 convolutional layers and each layer is followed by a Batch Normalization (BN) layer and a ReLU activate function. Then the features X and Y are fed into the CNN path and the Mamba path, respectively. In addition, a textual description of the image is sent to the text encoder of the pre-trained CLIP [26]. The result is a sequence of text tokens $T \in \mathbb{R}^{D\times L}$, where D and L represent the dimension and length of the text token sequence. This sequence is injected into the RDCB or MTB in each stage of the network.

Within a path of the encoder, the image features X or Y are first combined with the text tokens T using the TPB, to leverage the prompt effect of text. Then the resultant features are passed through the CFB, in which they are fused with the original image and the logit maps obtained from them, to incorporate prior characteristics and original image characteristics. Furthermore, the features are fed into the RDCB or MTB. An FFB bridges the RDCB and MTB at the same stage for the sake of performing feature fusion and interaction between the two paths. A downsampling operation is appended to each RDCB in the first three stages, which is implemented using a 3×3 convolutional layer with a stride of 2 and reduces the size of feature maps by half. The resultant feature maps of the two paths have dimensions of $8C \times H/8 \times W/8$.

On the other hand, there is not a CFB in the decoder. Regarding a RDCB or MTB contained in the second, third and fourth stages, the input feature maps are first upsampled using a 3×3 transposed convolutional layer with a stride of 2. These feature maps are then combined with those extracted using the RDCB or MTB at the symmetric stage in the encoder using a skip connection. The rest of the decoder mirror the symmetric components in the encoder. Finally, the outputs of the two paths are fused and processed using the classification block (see Fig. 2(f)) in order to obtain the segmentation result. This block comprises a linear layer and a 1×1 convolutional layer.

B. Text Prompt Block (TPB)

To exploit the semantic information of images, we propose a Text Prompt Block (TPB) (see Fig. 2(b)), which fuses the text tokens extracted from a textual description with the image features extracted at the CNN path or the Mamba path. Due to the TPB, the network is able to focus on the Region of Interest (ROI). As a result, the resultant features become more discriminant. Specifically, the corresponding textual description is sent to the text encoder of the CLIP [26] when an image is fed into the network. The output of the text encoder is a sequence of text tokens $T \in \mathbb{R}^{D \times L}$. Given a stage, these tokens are passed through a linear layer in order to adjust its dimensions.

With regard to a set of image feature maps $Z \in \mathbb{R}^{C \times H \times W}$, they are reshaped to $C \times L$, where $L = H \times W$. Then we fuse the image features and text tokens on the basis of the Multi-head Cross-Attention (MHCA) mechanism, which enables image features to actively retrieve textual semantics. Since different attention heads are able to capture the image-text relationships in different subspaces, MHCA establishes fine-grained correspondences from multiple perspectives. As a result, more precise image-text fusion can be achieved and the semantic understanding and object localization capabilities of the network can be enhanced.

The image feature maps Z are transformed into Q using a Query Linear layer, while the text tokens are processed using the Key and Value Linear layers to obtain K and V, respectively. The MHCA is computed by performing a matrix multiplication between Q and K, followed by a softmax function for normalization. The result is then multiplied by V and a set of fused features are produced, which enables text guidance in order to focus on relevant image regions. Additionally, the fused features are passed through a Linear layer to derive the more discriminant features. These features are reshaped back to the original shape. Finally, a residual connection is applied to those features for the purpose of preventing information loss during the fusion process. The above computation process can be expressed as

$$Q = \mathrm{W_Q}(\mathrm{Reshape}(Z)), \tag{1}$$

$$K = \mathrm{W_K}(T), \tag{2}$$

$$V = \mathrm{W_V}(T), \tag{3}$$

$$Z = \mathrm{Reshape}(\mathrm{Linear}(\mathrm{Softmax}(\frac{QK^T}{\sqrt{d_h}})V)) + Z, \tag{4}$$

where $\mathrm{W_Q}$, $\mathrm{W_K}$ and $\mathrm{W_V}$ represent the Query, Key and Value Linear layers, respectively and d_h denotes the number of cross-attention heads.

The use of the TPB at each stage of a path ensures that the text tokens which carry semantic information are injected into image features. In essence, the text tokens serve as a prompt for image feature extraction, which helps the network focus on the semantically important regions of the image.

C. Cascade Feature Block (CFB)

As shown in Fig. 2(c), we introduce a Cascade Feature Block (CFB), to incorporate prior characteristics and original image characteristics into the image representation. To be specific, the input feature maps are first passed through a 1×1 convolution to perform a simple decoding operation. The result is a set of logit maps. These maps are then concatenated with the original image. To fuse the logit maps with the image and restore the feature dimensions, they are sent to a second 1×1 convolution. The fused features are further added to the input feature maps. The computation can be expressed as

$$Logits = \mathrm{Conv}_{1\times1}(Z), \tag{5}$$

$$Z = \mathrm{Conv}_{1\times1}(\mathrm{Concat}(Logits, I)) + Z, \tag{6}$$

where Z represents the input feature maps, $Logits$ denotes the logit maps obtained by the decoding operation and I represents the original image.

As a result, the prior characteristics and original characteristics encoded in the logit maps and the original image, respectively, are injected into the image representation. This process enhances the discrimination ability of the representation.

D. Residual Double-Convolution Block (RDCB)

The CNN path of our TPCM-SegNet comprises a set of RDCBs, each of which consists of N Convolutional Sub-blocks (CB). The RDCB can be expressed as

$$\mathrm{RDCB}(X) = \mathrm{CB}^N(X), \tag{7}$$

where $\mathrm{CB}^N(X)$ means that X will be consecutively processed by the convolutional sub-block (CB) for N times. The CB, following a normal residual block, can be formulated as

$$\hat{X} = \mathrm{GELU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(X))), \tag{8}$$

$$X = \mathrm{GELU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\hat{X}))) + X, \tag{9}$$

where $GELU(\cdot)$ and $BN(\cdot)$ donote the Gaussian Error Linear Unit activation function and the Batch Normalization operation, respectively. Different numbers (N) of CBs are comprised of the RDCB at a stage of the network. Due to the inherent local inductive bias of CNNs, the CNN path extracts local features.

E. Mamba-Transformer Block (MTB)

The Mamba path of the TPCM-SegNet contains a series of MTBs. Each MTB consists of $N/2$ Mamba Sub-blocks

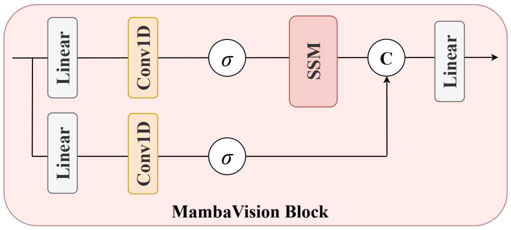

Fig. 3. The structure of the MambaVision [22] block. Here, the symbol σ represents the Sigmoid Linear Unit (SiLU) activation function.

(MB) and $N/2$ Transformer Sub-blocks (TB), which can be expressed as

$$\mathrm{MTB}(Y) = \mathrm{TB}^{N/2}(\mathrm{MB}^{N/2}(Y)), \tag{10}$$

where $\mathrm{MB}^{N/2}(Y)$ and $\mathrm{TB}^{N/2}(Y)$ indicate that Y will be consecutively processed by the MB and TB, respectively, for $N/2$ times. During this process, the shape of Y is first transformed from $C \times H \times W$ to $C \times L$, where $L = H \times W$, to match the computational dimensions. Y is then restored to its original shape after the processing has been completed. Similarly, different numbers (N) of MBs and TBs are used to construct the MTB at a stage of the network. The more comprehensive extraction of global features can be achieved by combining Mamba with Transformer to capture long-range dependencies.

With regard to the MB, the core component is the MambaVision Block (MVB) [22]. As shown in Fig. 3, the MVB contains a symmetric branch without an SSM [18], to compensate for the content loss caused by the sequential constraints of the SSM [18]. The two branches are then concatenated and fed into a linear layer. This process can be expressed as

$$Y_1 = \mathrm{SSM}(\sigma(\mathrm{Conv1D}(\mathrm{Linear}(Y)))), \tag{11}$$

$$Y_2 = \sigma(\mathrm{Conv1D}(\mathrm{Linear}(Y))), \tag{12}$$

$$Y = \mathrm{Linear}(\mathrm{Concat}(Y_1, Y_2)), \tag{13}$$

where $\sigma(\cdot)$ denotes the Sigmoid Linear Unit (SiLU) activation function. The above design ensures that the resultant feature representation incorporates both the sequential and spatial information, which exploits the strengths of both branches.

Additionally, the MB comprises two Layer Normalization (LN) layers, a Multi-layer Perceptron (MLP) unit and two residual connections, which further improves its representation capability. The MB can be formulated as:

$$\hat{Y} = \mathrm{MVB}(\mathrm{LN}(Y)) + Y, \tag{14}$$

$$Y = \mathrm{MLP}(\mathrm{LN}(\hat{Y})) + \hat{Y}. \tag{15}$$

The resultant feature maps Y are fed into the TB for further extraction of long-range dependencies. The structure of the TB is similar to that of the MB. But the MVB is replaced by the Multi-head Self-attention (MSA) [11] mechanism. The computation process of the TB can be expressed as:

$$\hat{Y} = \mathrm{MSA}(\mathrm{LN}(Y)) + Y, \tag{16}$$

$$Y = \mathrm{MLP}(\mathrm{LN}(\hat{Y})) + \hat{Y}. \tag{17}$$

Due to both the MB and TB, the MTB can effectively extract long-range dependencies from the image.

F. Feature Fusion Block (FFB)

As shown in Fig. 2(d), we design a Feature Fusion Block (FFB), which enables interaction and fusion between the features extracted at the CNN and Mamba paths. The FFB receives two sets of features extracted at the CNN and Mamba paths, respectively. To perform an initial fusion, these features are concatenated and fed into a 1×1 convolutional layer. The fused features are then added to the input features that require the injection operation. This process enables feature fusion while preserving the original features. The applications of the FFB to the CNN path and the Mamba path can be formulated as

$$X = \mathrm{Conv}_{1\times1}(\mathrm{Concat}(X, Y)) + X, \tag{18}$$

and

$$Y = \mathrm{Conv}_{1\times1}(\mathrm{Concat}(Y, X)) + Y, \tag{19}$$

respectively. As a result, the local and global features extracted using the two paths can be fused.

Given an RDCB and an MTB at the same stage, the intermediate feature maps $X^{(N/2)}$ produced by $N/2$ consecutive CBs and the feature maps Y generated by the prior CFB, are fed into an FFB. The resultant features are sent to the MTB rather than the features Y. This process incorporates the local features into the Mamba path, which can be expressed as

$$X^{(N/2)} = \mathrm{CB}^{N/2}(X), \tag{20}$$

$$Y = \mathrm{MTB}(\mathrm{FFB}(Y, X^{(N/2)})), \tag{21}$$

where $\mathrm{CB}^{N/2}(X)$ indicates that X is processed by $N/2$ consecutive CBs.

On the other hand, the two sets of features extracted using the two blocks, respectively, are fed into a second FFB. The fused features are used as the output of the RDCB instead of the features X. This operation injects the global features into the CNN paths, which can be expressed as

$$X = \mathrm{FFB}(\mathrm{RDCB}(X), Y). \tag{22}$$

Due to the above operations, information exchange between the two paths is achieved, which enables the fusion of local and global features and generates the more powerful features.

IV. Experimental Setup

In this section, we introduce the datasets, performance metrics, baselines and implementation details used in our experiments.

A. Datasets

To evaluate the effectiveness of our method in identifying organs, lesions and defects, we conducted a series of comparative experiments on eight datasets, including Synapse [42], ACDC [43], ISIC2017 [44], ISIC2018 [45], Kavsir-Seg [46], LIACi [47], CFD [48] and OUC-Crack [49]. Among these datasets, Synapse and ACDC are medical organ datasets, which contain 30 and 100 cases, respectively. Following the setup used in previous studies [12], [13], 18 and 12 cases in the Synapse dataset were used for training and testing, respectively, while 70, 10 and 20 cases in the ACDC dataset were utilized for training, validation and testing, respectively. The ISIC2017, ISIC2018 and Kavsir-Seg datasets are medical lesion datasets, which comprise 2,150, 2,694 and 1,000 images, respectively. In addition, the LIACi, CFD and OUC-Crack datasets are defect datasets, which consist of 1,561, 118 and 1,291 images, respectively. We split the ISIC2017 and ISIC2018 datasets into the training and testing sets at the ratio of 7:3. With regard to the Kavsir-Seg, LIACi, OUC-Crack and CFD datasets, they were divided into the training and testing sets using a 9:1 ratio.

B. Performance Metrics

We utilized the 95% Hausdorff Distance (HD95) and the Dice Similarity Coefficient (DSC) as performance metrics for the Synapse [42] dataset while only using DSC for the ACDC [43] dataset, following existing studies [12], [13]. For the remaining datasets, we used mean Intersection over Union (mIoU) and DSC as performance metrics. The HD95 metric can be formulated as

$$\begin{aligned} HD_{95} = \max \Big\{ & P_{95}\Big(\max_{x\in X}\min_{y\in Y}\|x-y\|\Big), \\ & P_{95}\Big(\max_{y\in Y}\min_{x\in X}\|y-x\|\Big)\Big\}, \end{aligned} \tag{23}$$

where X stands for the labels predicted, Y denotes ground-truth labels and $P_{95}(\cdot)$ represents the 95th percentile function. The DSC and mIoU metrics can be expressed as

$$DSC = \frac{2|X\cap Y|}{|X|+|Y|}, \tag{24}$$

$$mIoU = \frac{|X\cap Y|}{|X\cup Y|}. \tag{25}$$

C. Baselines

For comparison purposes, we used different baselines for the eight datasets. Regarding the Synapse [42] and ACDC [43] datasets, 27 and 23 baselines were utilized, respectively, following existing studies [16], [17]. For the other datasets, eight baselines [8], [9], [12], [13], [6], [4], [5], [50] were used.

D. Implementation Details

We resized the images to the resolution of 244×244 pixels. The training operation was conducted for 450 epochs with the batch size of 4. We used AdamW as the optimizer with the poly learning rate decay strategy. The learning rate, momentum and weight decay were set to 5e-5, 0.9 and 0.05, respectively. We employed the cross-entropy loss function. The CLIP [26] text encoder was initialized using the weights of the pre-trained ViT-B/32 [26], [35]. Our network was implemented using PyTorch 3.12. All experiments were performed on an NVIDIA GeForce RTX 3090 Graphics Processing Unit (GPU) with a memory usage of 22 GB.

V. Experimental Results

In this section, we first compare our TPCM-SegNet with different baselines on the eight publicly available datasets. To examine the impact of different components of the TPCM-SegNet, we also conducted a series of ablation experiments.

A. Comparative Experiments

1) Evaluation on Synapse and ACDC: To assess the segmentation performance of the TPCM-SegNet on organ structures and potential anomalies in medical images, we tested our method on the Synapse [42] and ACDC [43] datasets, together with 27 and 23 baselines, respectively. The results of the quantitative comparison are presented in Tables I and II.

With regard to the Synapse dataset, our method achieved the better result, compared to the 27 baselines, in terms of two average metrics. It should be noted that our method produced the average HD95 value of 6.83, which exceeded the values obtained using the baselines with large margins. In terms of the average DSC metric, our method outperformed AgileFormer [16] by 1.14%, which ranked the second. Compared with ICCT-UNet [6], our method adopted a combination of Mamba and Transformer in the global feature extraction path, which enabled the extraction of the richer global features than those extracted using the pure Transformer path of ICCT-UNet.

In addition, our method incorporated textual information and prior characteristics, which further enhanced the representation capability of the network. As a result, our approach achieved improvements of 2.28% and 4.18 in terms of the DSC and HD95 metrics, respectively, over ICCT-UNet. To be exact, the proposed TPCM-SegNet derived the average DSC value of 86.88%, which outperformed the 27 baselines. When only a single category was considered, our method produced the best result on the Aorta, Gallbladder, Kidney (R), Liver, Pancreas and Stomach categories and achieved the comparable performance on the Kidney (L) and Spleen categories.

When the ACDC dataset was used, our method achieved the average DSC value of 92.70%, which outperformed all 23 baselines. In addition, the TPCM-SegNet produced the best result on the Left Ventricle (LV) category and the second-best result on the Right Ventricle (RV) category.

As shown in Fig. 4, we compared the results of our method and five baselines obtained on the Synapse [42] and ACDC [43] datasets. It can be observed that LeVIT-UNet [51], MT-UNet [52], TransUNet [12], SwinUnet [13] and ICCTUNet [6] exhibited noticeable mis-segmentation issues on both datasets. For example, LeVIT-UNet [51], TransUNet [12] and ICCTUNet [6] misidentified large areas of the spleen as the liver. In contrast, our method correctly segmented these cases while maintaining the higher degree of segmentation integrity. However, our TPCM-SegNet produced minor deficiencies in detecting the Kidney (L) and Spleen categories, as shown in Fig. 5. These deficiencies might result from the inherent bias of our method, such as a lack of ability in capturing texture or boundary characteristics of the Kidney (L) and Spleen categories, leading to relatively insufficient activations. Those deficiencies could be mitigated by applying data augmentation

TABLE I

Comparison between 27 baselines and the proposed TPCM-SegNet on the Synapse [42] dataset. For each metric, the top three best results are highlighted in **RED**, *CYAN* and BLUE Fonts, respectively.

Network	Avg.		DSC ↑							
	DSC ↑	HD95 ↓	Aotra	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
R50 U-Net [12]	74.68	36.87	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
R50 Att-Unet [12]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
LeViT-Unet [51]	78.53	16.84	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
TransUNet [12]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUnet [13]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MT-UNet [52]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MISSFormer [36]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
ScaleFormer [53]	82.86	16.81	88.73	74.97	86.36	83.31	95.12	64.85	89.40	80.14
HiFormer [54]	80.39	14.70	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
SSFormerPVT [55]	78.01	25.72	82.78	63.74	80.72	78.11	78.11	61.53	87.07	76.61
TransCeption [56]	82.24	20.89	87.60	71.82	86.23	80.29	95.01	65.27	91.68	80.02
SSTrans-Net [57]	82.89	15.55	86.72	72.87	86.99	83.33	94.61	64.98	92.23	81.37
CASTformer [58]	82.55	22.73	89.05	67.48	86.05	82.17	95.61	67.49	91.00	81.55
PVT-CASCADE [59]	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
TransCASCADE [59]	82.68	17.34	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52
PAG-TransYnet [60]	83.43	15.82	89.67	68.89	86.74	84.88	*95.87*	68.75	92.01	80.66
ParaTransCNN [61]	83.86	15.86	88.12	68.97	87.99	83.84	95.01	69.79	**92.71**	84.43
PVT-EMCAD-B0 [31]	81.97	17.39	87.21	66.62	87.48	83.96	94.57	62.00	*92.66*	81.22
PVT-EMCAD-B2 [31]	83.63	15.68	88.14	68.87	88.08	84.10	95.26	68.51	92.17	83.92
VM-UNet [4]	81.08	19.21	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40
ICCT-UNet [6]	84.60	*11.01*	*91.13*	71.50	86.58	83.86	95.64	*72.20*	91.81	84.07
Parallel MERIT [17]	84.22	16.51	88.38	73.48	87.21	84.31	95.06	69.97	91.21	84.15
Cascaded MERIT [17]	84.90	13.22	87.71	74.40	87.79	84.85	95.26	71.81	92.01	85.38
AgileFormer [16]	*85.74*	18.70	89.11	*77.89*	*88.83*	*85.00*	95.64	71.62	92.20	*85.63*
CoTransUNet [62]	82.39	17.51	87.13	71.34	86.40	81.12	94.88	66.27	90.37	81.63
MCAFT [63]	83.87	14.20	88.56	74.20	86.11	84.76	95.33	69.06	91.08	81.87
kMaXU [64]	84.31	16.58	88.10	68.67	**89.29**	84.42	95.48	72.74	92.53	83.22
TPCM-SegNet (Ours)	**86.88**	**6.83**	**91.18**	**78.20**	87.71	**85.74**	**95.97**	**75.68**	92.55	**88.03**

TABLE II

Comparison between the proposed TPCM-SegNet and 23 baselines on the ACDC [43] dataset.

Network	Avg. DSC ↑	RV	Myo	LV
R50 U-Net [12]	87.55	87.10	80.63	94.92
R50 Att-UNet [12]	86.75	87.58	79.20	93.47
LeViT-UNet [51]	90.32	89.55	87.64	93.76
TransUNet [12]	89.71	88.86	84.53	95.73
SwinUnet [13]	90.00	88.55	85.62	95.83
MT-UNet [52]	90.43	86.64	89.04	95.62
MISSFormer [36]	90.86	89.55	88.04	94.99
ScaleFormer [53]	90.17	87.33	88.16	95.04
HiFormer [54]	90.12	91.06	84.54	94.77
TransCeption [56]	88.47	87.88	82.87	94.66
SSTrans-Net [57]	90.31	89.02	87.51	94.40
PVT-CASCADE [59]	91.46	89.97	88.90	95.50
TransCASCADE [59]	91.63	90.25	89.14	95.50
ParaTransCNN [61]	91.31	**92.76**	85.84	95.34
PVT-EMCAD-B0 [31]	91.34	89.37	88.99	95.65
PVT-EMCAD-B2 [31]	92.12	90.65	89.68	96.02
ICCT-UNet [6]	91.64	90.66	88.94	95.30
Parallel MERIT [17]	92.32	90.87	*90.00*	96.08
Cascaded MERIT [17]	91.85	90.23	89.53	95.80
AgileFormer [16]	*92.55*	91.05	**90.40**	*96.19*
CoTransUNet [62]	89.08	91.79	84.47	90.98
MCAFT [63]	92.32	91.07	89.87	96.05
kMaXU [64]	92.13	90.76	89.70	95.93
TPCM-SegNet (Ours)	**92.70**	*92.22*	89.67	**96.21**

or generative approaches to enhance the representation of underrepresented organs.

2) Evaluation on ISIC2017, ISIC2018 and Kavsir-Seg: We also compared our method with seven baselines on three anomaly lesion datasets, including ISIC2017 [44], ISIC2018 [45] and Kavsir-Seg [46]. As reported in Table III, the proposed method outperformed the seven baselines across the three datasets, regardless of which performance metric was used. To be specific, our method exceeded HMT-UNet [5] by 0.58% in terms of the mIoU metric on the ISIC2017 dataset. Given that the ISIC2018 dataset was utilized, our method achieved a DSC value higher than ICCTUNet [6] by 0.35%. It outperformed VMUNet [4] with a 0.78% increase in mIoU and a 0.26% increase in DSC on the Kavsir-Seg dataset.

In Fig. 6, (a-c) show the visualizations of the results and the corresponding zoomed-in views produced by our method and three baselines on the three lesion datasets, respectively. It can be seen that both TransUNet [12] and VMUNet [4] produced inaccurate results by misidentifying normal areas as anomaly lesion regions on an ISIC2017 image. When an ISIC2018 image was used, our method achieved the more accurate segmentation of the lesion area, compared to the three baselines. Regarding an image in the Kavsir-Seg dataset, our method successfully segmented all lesion regions, while the three baselines either missed one lesion or merged them into a single region.

3) Evaluation on LIACi, OUC-Crack and DeepCrack: Furthermore, we compared our method with seven baselines on three defect datasets, including LIACi [47], CFD [48] and OUC-Crack [49], as shown in Table IV. Given the underwater hull defect dataset, i.e., LIACi, our method outperformed its counterparts greatly. In contrast to XNet [9], our method achieved the margins of 11.41% and 10.34% in terms of the

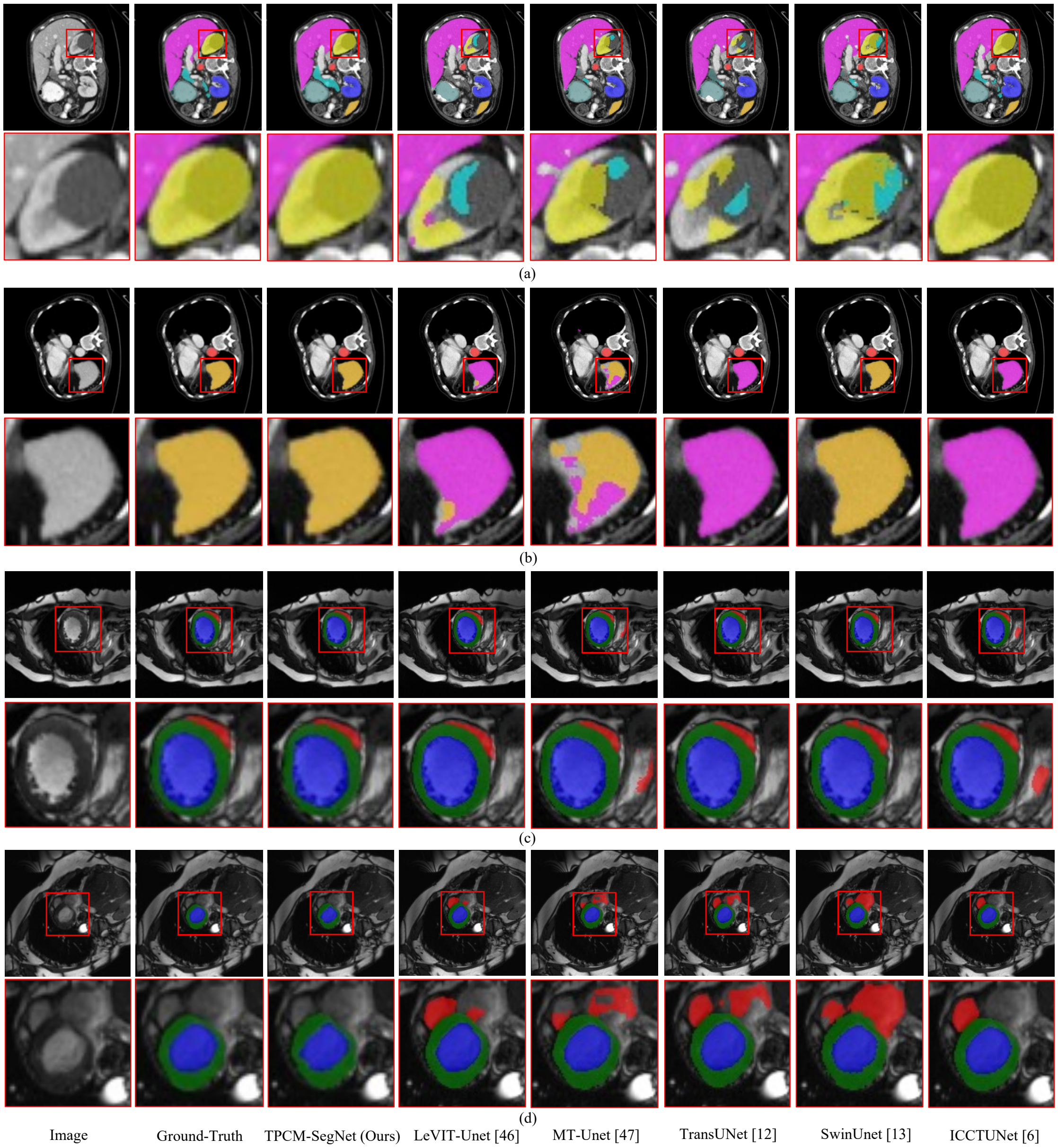

Fig. 4. The visualizations of the segmentation results and the corresponding zoomed-in views produced by our TPCM-SegNet and five baselines are presented. Here, (a) and (b) show the images obtained from the Synapse [42] dataset, while (c) and (d) present the images derived from the ACDC [43] dataset.

TABLE III
COMPARISON BETWEEN THE TPCM-SEGNET AND EIGHT BASELINES ON THE ISIC2017 [44], ISIC2018 [45] AND KAVSIR-SEG [46] DATASETS.

Network	ISIC2017		ISIC2018		Kavsir-Seg	
	mIoU ↑	DSC ↑	mIoU ↑	DSC ↑	mIoU ↑	DSC ↑
UNet [8]	77.02	84.55	78.64	87.22	78.44	87.24
XNet [9]	77.09	*84.95*	78.84	87.21	71.24	82.81
TransUNet [12]	76.78	83.07	79.00	87.01	75.18	84.80
SwinUnet [13]	75.79	83.27	78.24	86.78	62.16	72.34
ICCTUNet [6]	77.49	84.07	78.82	*87.44*	79.87	88.26
VMUNet [4]	76.77	83.95	79.06	86.74	*80.04*	*88.82*
HMT-UNet [5]	*77.72*	84.70	*79.37*	87.40	79.48	87.35
U-KAN [50]	76.47	84.41	79.28	87.20	73.50	83.04
TPCM-SegNet (Ours)	**78.30**	**85.07**	**79.38**	**87.79**	**80.82**	**89.06**

mIoU and DSC metrics, respectively. It is demonstrated that our method retained good performance even in challenging underwater environments. Using the pavement crack dataset, namely, CFD, the TPCM-SegNet also produced the superior results, compared to the seven baselines. Specifically, these results exceeded those derived using ICCTUNet [6] by 7.53% in mIoU and those obtained using XNet [9] by 7.77% in DSC. Our method achieved the best performance on the OUC-Crack wall crack dataset, which outperformed U-KAN [50] by 0.49% and 0.36% in terms of the mIoU and DSC metrics, respectively.

The visualization of the results and the associated zoomed-in views obtained using our method and three baselines on the three defect datasets are presented in (d-f) of Fig. 6, respectively. As can be seen, our method not only accurately located defects but also properly segmented the underwater hull structure contained in an LIACi image, which overcame the interference from the complex underwater environment. When the two crack datasets were used, our method produced

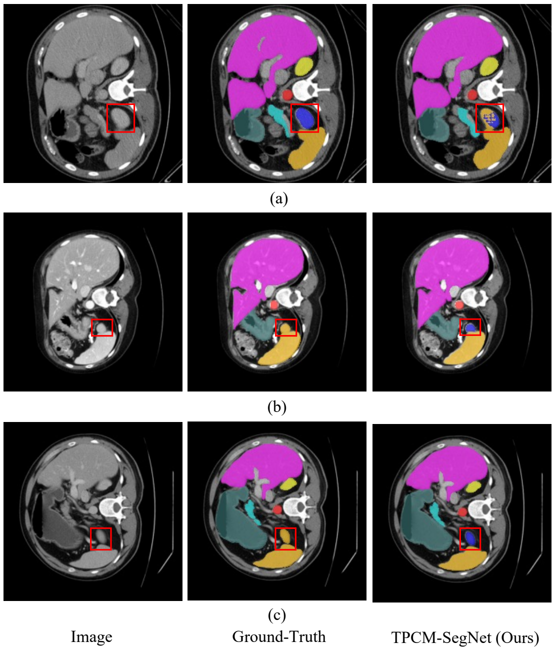

Fig. 5. Examples of segmentation deficiencies produced by TPCM-SegNet in terms of the Kidney (L) and Spleen categories.

TABLE IV
COMPARISON BETWEEN OUR TPCM-SEGNET AND EIGHT BASELINES ON THE LIACI [47], CFD [48] AND OUC-CRACK [49] DATASETS.

Network	LIACi		CFD		OUC-Crack	
	mIoU ↑	DSC ↑	mIoU ↑	DSC ↑	mIoU ↑	DSC ↑
UNet [8]	47.03	60.22	21.21	34.32	*57.85*	72.42
XNet [9]	*48.85*	*62.31*	*22.17*	*36.19*	56.80	71.83
TransUNet [12]	43.25	57.47	12.54	21.56	56.59	71.71
SwinUnet [13]	43.16	57.14	21.17	34.42	55.31	70.41
ICCTUNet [6]	47.42	60.70	21.32	35.09	57.37	72.39
VMUNet [4]	46.26	60.14	20.33	32.95	54.03	69.30
HMT-UNet [5]	43.35	56.57	21.09	34.23	54.82	69.66
U-KAN [50]	47.35	60.88	12.74	21.42	57.81	*72.68*
TPCM-SegNet (Ours)	**60.26**	**72.65**	**28.85**	**43.96**	**58.30**	**73.04**

the more adequate segmentation of cracks, which manifest fewer omissions and discontinuities, compared to its counterparts.

B. Ablation Studies

To investigate the effectiveness of different components of the proposed TPCM-SegNet, we conducted a series of comprehensive ablation experiments. For simplicity, only the Synapse [42] dataset was used.

1) Effect of the Dual-path Architecture: To examine the effectiveness of the dual-path architecture, we performed a comparative experiment using the single-path and dual-path architectures on top of convolution, Transformer, Mamba and Mamba-Transformer. The comparative results are shown in Table V.

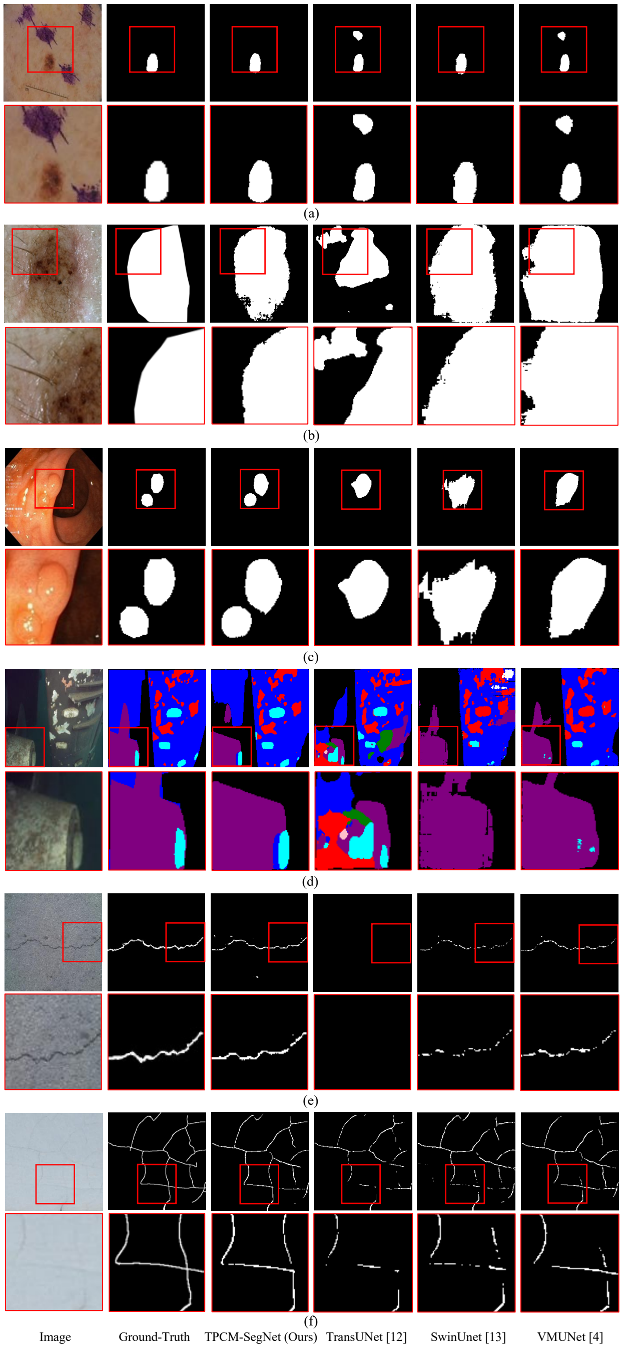

Fig. 6. The visualizations of the results and the associated zoomed-in views produced by our TPCM-SegNet and three baselines on six images contained in the ISIC2017 [44], ISIC2018 [45], Kavsir-Seg [46], LIACi [47], CFD [48] and OUC-Crack [49] datasets in turn.

Given that the single-path architecture was used, the feature fusion module was removed. As can be seen, the single-path convolutional network outperformed its three single-path counterparts. This should be due to the inherent local feature representation advantages of CNNs over the Transformer and/or Mamba networks. Within the dual-path architecture, one path captures local characteristics using a CNN, while the other path extracts global features using a Transformer network, a Mamba network, or a Mamba-Transformer network. As reported in Table V, the dual-path architecture greatly improved the performance of the model, compared to the single-path approach, because of the combination of the local and global features extracted at the two paths, respectively.

2) Effect of the Depth of Different Stages: As displayed in Fig. 2(a), our TPCM-SegNet consists of four symmetric stages, in which the encoding and decoding stages mirror each other. Both the RDCB and MTB in each stage comprise N

TABLE V
COMPARISON BETWEEN THE SINGLE-PATH AND DUAL-PATH ARCHITECTURES ALONG WITH DIFFERENT NETWORKS. HERE, C, T, M AND M-T REPRESENT CONVOLUTION, TRANSFORMER, MAMBA AND MAMBA-TRANSFORMER, RESPECTIVELY.

Architecture	C	T	M	M-T	DSC ↑	HD95 ↓
Single-Path	✓	-	-	-	84.24	10.35
Single-Path	-	✓	-	-	78.53	15.26
Single-Path	-	-	✓	-	76.15	24.19
Single-Path	-	-	-	✓	79.56	17.73
Dual-Path	✓	✓	-	-	86.40	9.95
Dual-Path	✓	-	✓	-	86.69	6.87
Dual-Path	✓	-	-	✓	**86.88**	**6.83**

TABLE VI
EFFECT OF THE COMBINATION OF DIFFERENT N VALUES USED IN FOUR STAGES ON THE PERFORMANCE OF OUR TPCM-SEGNET. HERE, ΔDSC AND ΔHD95 WERE COMPUTED BY COMPARING WITH THE TWO VALUES OBTAINED USING THE FIRST COMBINATION, RESPECTIVELY.

Combination	DSC ↑	HD95 ↓	ΔDSC	ΔHD95
[2, 2, 2, 2]	86.09	8.43	-	-
[2, 2, 2, 4]	86.35	8.13	+0.26	-0.30
[2, 2, 4, 4]	86.29	8.05	+0.20	-0.83
[2, 2, 4, 8]	85.88	9.83	-0.21	+1.4
[2, 4, 4, 8]	**86.88**	**6.83**	**+0.79**	**-1.6**
[2, 4, 8, 8]	85.01	11.65	-1.08	+3.22
[2, 4, 8, 16]	84.75	10.91	-1.34	+2.48

sub-blocks. The value of N determines the depth of a stage. To investigate the impact of the depth of different stages on the performance of the TPCM-SegNet, we conducted a series of experiments. As shown in Table VI, the combination of the four N values was initialized as $[2, 2, 2, 2]$ and we gradually increased the N value used in each stage. It can be observed that the performance of our TPCM-SegNet was improved by increasing the N value at the beginning, while a performance degradation was caused when the value of N exceeded a certain number. Among the combinations that we tested, $[2, 4, 4, 8]$ produced the optimal results.

3) Comparison of Different Feature Fusion Methods: Within each stage of the TPCM-SegNet, we used two FFBs to perform interaction and fusion between the features extracted at the CNN and Mamba paths. To examine the effectiveness of the FFB, we compared the TPCM-SegNet with its three variants, which were derived by removing FFBs from it, replacing the FFB by an addition operation and replacing the FFB by a concatenation operation, respectively. The results obtained using the four networks are shown in Fig. 7. It can be seen that the DSC and HD95 values obtained without feature fusion are 85.28% and 8.92, respectively. Surprisingly, the two values derived using the addition and concatenation variants were worse than these values. It is indicated that the two simple feature fusion methods failed to effectively integrate local and global features. In contrast, the DSC and HD95 values obtained using the TPCM-SegNet, together with the FFB, are 86.88% and 6.83, respectively. It is demonstrated that the FFB benefits the performance of the proposed TPCM-SegNet via effectively fulfilling feature interaction and fusion between the two paths.

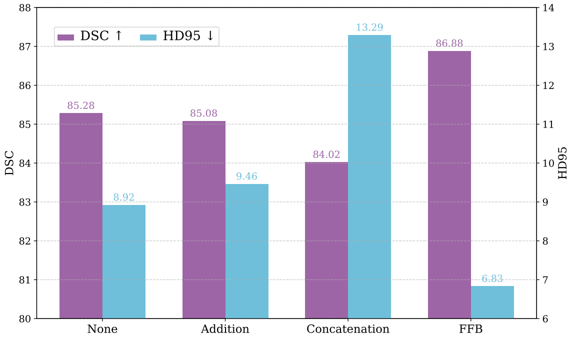

Fig. 7. Comparison of different feature fusion methods used for the TPCM-SegNet.

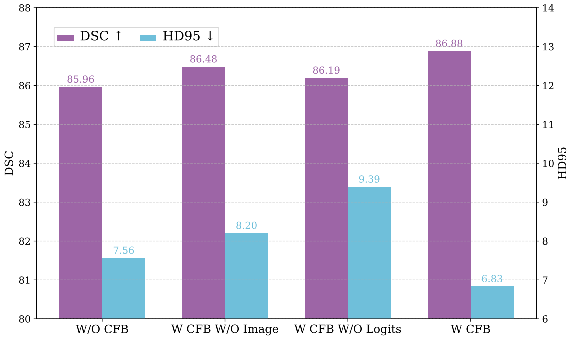

Fig. 8. Comparison between the TPCM-SegNet and its three variants obtained by removing the Cascade Feature Block (CFB) or removing the information used by the CFB.

4) Effect of the Cascade Feature Block: We utilized the Cascade Feature Block (CFB) in the encoding stages for the sake of integrating prior characteristics and original image characteristics. To verify the usefulness of the CFB, we compared the TPCM-SegNet with its three variants. These variants were derived by removing the CFB from the TPCM-SegNet, removing the original image information from the CFB and removing the logit maps from the CFB. The comparative results are shown in Fig. 8. As can be seen, the two variants, which used the CFB but discarded the original image or logit maps, produced the higher DSC values, compared to the variant which removed the CFD. In this case, the HD95 value also increased, which indicates the worse performance. However, the application of the CFB to our TPCM-SegNet produced the best result in terms of both metrics. It is suggested that the introduction of the CFB is useful for improving the performance of our TPCM-SegNet.

5) Effect of the Text Prompt Block: To investigate the effect of different image–text fusion strategies on our network, we applied them between the encoder and decoder of TPCM-SegNet. As shown in Table VII, both the addition and concatenation operations improved the segmentation accuracy. However, the fusion operation adopted on top of cosine similarity performed poorly, which resulted in a decrease in performance. In contrast, the proposed Text Prompt Block (TPB) achieved the best performance with the largest performance gain.

To further verify the effectiveness of the TPB, we performed an additional ablation experiment by progressively incorporat-

TABLE VII
COMPARISON OF DIFFERENT IMAGE-TEXT FUSION STRATEGIES WHEN USED TOGETHER WITH TPCM-SEGNET.

Fusion Strategy	DSC ↑	HD95 ↓	ΔDSC	ΔHD95
None	83.21	13.66	-	-
Addition	84.28	10.82	+1.07	-2.84
Concatenation	84.42	9.34	+1.21	-4.32
Cosine Similarity	81.78	12.44	-1.43	-1.22
Text Prompt Block (TPB)	**84.97**	**8.62**	**+1.76**	**-5.04**

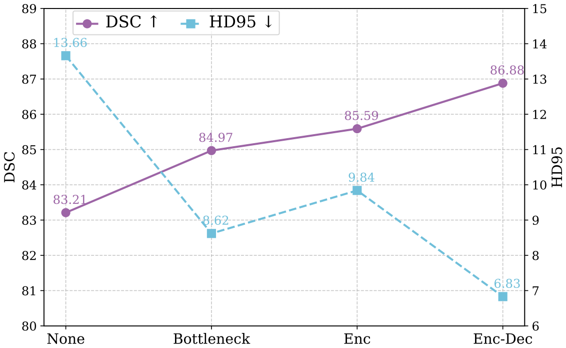

Fig. 9. Comparison of the variants of the TPCM-SegNet obtained by incorporating text tokens through the TPB at different positions of the network.

TABLE VIII
COMPARISON OF DIFFERENT COMBINATIONS OF THE BLOCKS IN TPCM-SEGNET IN TERMS OF THE NUMBER OF PARAMETERS, COMPUTATIONAL COMPLEXITY (FLOPS), DSC AND HD95.

Combination	Param. (M) ↓	FLOPs (G) ↓	DSC ↑	HD95 ↓
RDCB&MTB	**147.8**	**246.2**	81.15	16.78
RDCB&MTB+FFB	153.1	255.6	82.99	13.98
RDCB&MTB+FFB+CFB	153.2	255.7	83.21	13.66
RDCB&MTB+FFB+CFB+TPB&CLIP	183.4	264.3	**86.88**	**6.83**

TABLE IX
COMPARISON BETWEEN NINE STATE-OF-THE-ART NETWORKS WITH THE PROPOSED TPCM-SEGNET IN TERMS OF THE NUMBER OF PARAMETERS, COMPUTATIONAL COMPLEXITY (FLOPS), DSC AND HD95.

Network	Param. (M) ↓	FLOPs (G) ↓	DSC ↑	HD95 ↓
TransUNet [12]	101.1	28.5	77.48	31.69
SwinUnet [13]	41.3	8.7	79.13	21.55
MT-UNet [52]	79.1	44.8	78.59	26.59
HiFormer [54]	**25.5**	17.8	80.39	14.70
MissFormer [36]	35.4	**7.2**	81.96	18.20
ScaleFormer [53]	113.7	48.5	82.86	16.81
ICCT-UNet [6]	67.4	90.9	84.60	11.01
Cascade MERIT [17]	147.9	33.3	84.90	13.22
AgileFormer [16]	112.6	24.9	85.74	18.70
TPCM-SegNet (Ours)	183.4	264.3	**86.88**	**6.83**

ing text tokens through the TPB at different positions of the TPCM-SegNet, including the encoder, bottleneck[2] and both the encoder and decoder. As shown in Fig. 9, the application of the TPB to different positions normally produced the better results than those obtained by removing the TPBs from the TPCM-SegNet. To be specific, the best DSC value 86.88% was derived by applying the TPB to both the encoder and decoder, which exceeded the value of 83.21% obtained without using the TPB by a large margin. In terms of the HD95 metric, the best performance was also achieved by applying the TPB to both the encoder and decoder, which was much better than that obtained without using the TPB. These results demonstrate that the TPB is able to boost the performance of our network by incorporating semantic information into the feature representation.

6) Comparison of Different Combinations of the Blocks in TPCM-SegNet: To evaluate the computational overhead required by different blocks in TPCM-SegNet, we conducted a comparative experiment. Specifically, we used a dual-path network, which comprised RDCBs and MTBs, as a baseline. Then we progressively incorporated additional blocks into the network to derive a new variant. As reported in Table VIII, the number of parameters and FLOPs increased by 5.3M and 9.4G, respectively, after the FFBs had been incorporated, while the DSC and HD95 values increased by 1.84% and 2.8, respectively. With the further incorporation of the CFBs, only a slight computational overhead was introduced, resulting in improvements of 0.22% and 0.32 in DSC and HD95, respectively. Finally, the number of parameters and FLOPs increased by 30.2M and 8.6G, respectively, when the text encoder of the CLIP and the TPB were used. However, great performance gains were also achieved, in which the DSC and HD95 values were improved by 3.67% and 6.83, respectively.

In addition, we compared our TPCM-SegNet with nine image segmentation networks in terms of the number of parameters, computational complexity (FLOPs), DSC and HD95. As shown in Table IX, our network involves more parameters and higher FLOPs, compared to its counterparts. This should be attributed to the dual-path architecture and the incorporation of text tokens. However, our TPCM-SegNet achieves the better segmentation performance, with large margins in both the DSC and HD95 metrics.

7) Effect of Different Loss Functions: We investigated the impact of different loss functions on the performance of our TPCM-SegNet. As shown in Table X, the Cross-Entropy Loss that we used achieved the best result in terms of each metric. In particular, the DSC and HD95 values were improved by 4.32% and 8.71, respectively, compared with the Binary Cross-Entropy Loss. In addition, both the Focal Loss and Dice Loss, which were normally used to address the class imbalance issue, were outperformed by the Cross-Entropy Loss.

8) Effect of Skip Connections: To examine the effect of skip connections on the proposed TPCM-SegNet, we applied them to different paths of the network. As reported in Table XI, the application to the CNN or Mamba path normally produced the better result than that derived without using skip connections. However, the best result was obtained by applying skip connections to both paths, which outperformed the network without using skip connections by 3.06% and 3.27 in terms of DSC and HD95, respectively. It is suggested that skip connections play an important role in our TPCM-SegNet.

VI. CONCLUSION

In this study, we introduced a Text-Prompted Dual-Path Convolution-Mamba Network for anomaly segmentation, namely, TPCM-SegNet, to address the limitations of existing methods in extracting discriminant image features and utilizing semantic information. The TPCM-SegNet consisted of two parallel paths, which aimed to extract local and global features

[2] Here, it is meant that a TPB was added between the encoder and decoder.

TABLE X
Effect of different loss functions on the performance of our TPCM-SegNet.

Loss Function	DSC ↑	HD95 ↓
Dice Loss	85.37	8.65
Focal Loss	84.59	12.78
Cross-Entropy Loss	**86.88**	**6.83**
Binary Cross-Entropy Loss	82.56	15.54

TABLE XI
Comparison of the skip connections applied to different paths of our TPCM-SegNet.

Skip Connections		DSC ↑	HD95 ↓	ΔDSC	ΔHD95
CNN Path	Mamba Path				
-	-	83.82	10.10	-	-
✓	-	84.33	10.13	+0.45	+0.03
-	✓	84.60	8.33	+0.78	-1.77
✓	✓	**86.88**	**6.83**	**+3.06**	**-3.27**

and were built on top of Residual Double-Convolution Blocks (RDCBs) and Mamba-Transformer Blocks (MTBs), respectively. To enable the interaction and fusion of the features extracted using the RDCB and MTB at the same stage, we proposed a Feature Fusion Block (FFB). Considering the importance of semantic information to image analysis, we further designed a Text Prompt Block (TPB), which fused the text tokens extracted from a textual description with image features. As a result, the semantics understanding ability of the network was improved. In addition, we adopted a Cascade Feature Block (CFB), which used a cascading mechanism to combine the feature maps, the logit maps decoded from them and the original image in each stage of the encoder. In essence, the CFB injected prior and original characteristics into the image representation. Experimental results showed that the proposed TPCM-SegNet produced the better, or at least competitive, results, compared to different baselines, across eight publicly available datasets. We believe that these results should be due to the strong image representation and semantic understanding capability of our TPCM-SegNet.

However, both the dual-path architecture and image-text feature fusion strategy inevitably increase the computational cost of our method. In our future work, we will manage to improve segmentation accuracy through a multimodal fusion strategy, while maintaining a lightweight network design.

References

[1] J. Cen, Z. Jiang, L. Xie, D. Jiang, W. Shen, and Q. Tian, "Consensus synergizes with memory: A simple approach for anomaly segmentation in urban scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 1086–1097, 2023.

[2] P. Xing, Y. Sun, D. Zeng, and Z. Li, "Normal image guided segmentation framework for unsupervised anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 4639–4652, 2024.

[3] J. Dong, Y. Cong, G. Sun, Y. Yang, X. Xu, and Z. Ding, "Weakly-supervised cross-domain adaptation for endoscopic lesions segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 2020–2033, 2021.

[4] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2402.02491*, 2024.

[5] M. Zhang, Z. Chen, Y. Ge, and X. Tao, "Hmt-unet: a hybird mamba-transformer vision unet for medical image segmentation," *arXiv preprint arXiv:2408.11289*, 2024.

[6] H. Qi, H. Zhou, J. Dong, and X. Dong, "Small sample image segmentation by coupling convolutions and transformers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5282–5294, 2023.

[7] H. Yao, W. Yu, W. Luo, Z. Qiang, D. Luo, and X. Zhang, "Learning global-local correspondence with semantic bottleneck for logical anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3589–3605, 2024.

[8] 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: 18th international conference, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[9] J. Bullock, C. Cuesta-Lázaro, and A. Quera-Bofarull, "Xnet: a convolutional neural network (cnn) implementation for medical x-ray image segmentation suitable for small datasets," in *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10953. SPIE, 2019, pp. 453–463.

[10] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "Deepcrack: Learning hierarchical convolutional features for crack detection," *IEEE transactions on image processing*, vol. 28, no. 3, pp. 1498–1512, 2018.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[14] Y. Chen, C. Zhang, B. Chen, Y. Huang, Y. Sun, C. Wang, X. Fu, Y. Dai, F. Qin, Y. Peng *et al.*, "Accurate leukocyte detection based on deformable-detr and multi-level feature fusion for aiding diagnosis of blood diseases," *Computers in biology and medicine*, vol. 170, p. 107917, 2024.

[15] Y. Chen, Z. Zhu, S. Zhu, L. Qiu, B. Zou, F. Jia, Y. Zhu, C. Zhang, Z. Fang, F. Qin *et al.*, "Sckansformer: Fine-grained classification of bone marrow cells via kansformer backbone and hierarchical attention mechanisms," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[16] P. Qiu, J. Yang, S. Kumar, S. S. Ghosh, and A. Sotiras, "Agileformer: spatially agile transformer unet for medical image segmentation," *arXiv preprint arXiv:2404.00122*, 2024.

[17] M. M. Rahman and R. Marculescu, "Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1526–1544.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[19] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2025.

[20] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[21] J. Xu, "Hc-mamba: Vision mamba with hybrid convolutional techniques for medical image segmentation," *arXiv preprint arXiv:2405.05007*, 2024.

[22] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," *arXiv preprint arXiv:2407.08083*, 2024.

[23] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," in *European Conference on Computer Vision*. Springer, 2025, pp. 74–91.

[24] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "Towards language-free training for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 907–17 917.

[25] J. Jain, A. Singh, N. Orlov, Z. Huang, J. Li, S. Walton, and H. Shi, "Semask: Semantically masked transformers for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 752–761.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable

visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[27] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "San: side adapter network for open-vocabulary semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 546–15 561, 2023.

[28] B. Xie, J. Cao, J. Xie, F. S. Khan, and Y. Pang, "Sed: A simple encoder-decoder for open-vocabulary semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 3426–3436.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] Z. Zhang, R. Peng, Y. Hu, and R. Wang, "Geomvsnet: Learning multi-view stereo with geometry perception," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 508–21 518.

[31] M. M. Rahman, M. Munir, and R. Marculescu, "Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 769–11 779.

[32] Y. Chen, B. Zou, Z. Guo, Y. Huang, Y. Huang, F. Qin, Q. Li, and C. Wang, "Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 7759–7767.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[34] H. Li, D.-H. Zhai, and Y. Xia, "Erdunet: An efficient residual double-coding unet for medical image segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2083–2096, 2023.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[36] X. Huang, Z. Deng, D. Li, and X. Yuan, "Missformer: An effective medical image segmentation transformer," *arXiv preprint arXiv:2109.07162*, 2021.

[37] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[38] J. Wang, J. Chen, D. Chen, and J. Wu, "Large window-based mamba unet for medical image segmentation: Beyond convolution and self-attention," *arXiv e-prints*, pp. arXiv–2403, 2024.

[39] Z. Jin, B. Liu, Q. Chu, and N. Yu, "Isnet: Integrate image-level and semantic-level context for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7189–7198.

[40] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 134–18 144.

[41] J.-J. Wu, A. C.-H. Chang, C.-Y. Chuang, C.-P. Chen, Y.-L. Liu, M.-H. Chen, H.-N. Hu, Y.-Y. Chuang, and Y.-Y. Lin, "Image-text co-decomposition for text-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 794–26 803.

[42] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, vol. 5. Munich, Germany, 2015, p. 12.

[43] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[44] M. Berseth, "Isic 2017-skin lesion analysis towards melanoma detection," *arXiv preprint arXiv:1703.00523*, 2017.

[45] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[46] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia modeling: 26th international conference, MMM 2020, 2020, proceedings, part II 26*. Springer, 2020, pp. 451–462.

[47] M. Waszak, A. Cardaillac, B. Elvesæter, F. Rødølen, and M. Ludvigsen, "Semantic segmentation in underwater ship inspections: Benchmark and data set," *IEEE Journal of Oceanic Engineering*, vol. 48, no. 2, pp. 462–473, 2022.

[48] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3434–3445, 2016.

[49] B. Xu, W. Shao, and X. Dong, "Drone-based wall crack detection using model-agnostic meta-learning," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 15 116–15 128, 2025.

[50] C. Li, X. Liu, W. Li, C. Wang, H. Liu, Y. Liu, Z. Chen, and Y. Yuan, "U-kan makes strong backbone for medical image segmentation and generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, 2025, pp. 4652–4660.

[51] G. Xu, X. Zhang, X. He, and X. Wu, "Levit-unet: Make faster encoders with transformer for medical image segmentation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2023, pp. 42–53.

[52] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Mixed transformer u-net for medical image segmentation," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 2390–2394.

[53] H. Huang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Scaleformer: Revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation," in *IJCAI*, 2022.

[54] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 6202–6212.

[55] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise feature fusion: Local guides global," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 110–120.

[56] R. Azad, Y. Jia, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Enhancing medical image segmentation with transception: A multi-scale feature fusion approach," *arXiv preprint arXiv:2301.10847*, 2023.

[57] L. Fu, Y. Chen, W. Ji, and F. Yang, "Sstrans-net: smart swin transformer network for medical image segmentation," *Biomedical Signal Processing and Control*, vol. 91, p. 106071, 2024.

[58] C. You, R. Zhao, F. Liu, S. Dong, S. Chinchali, U. Topcu, L. Staib, and J. Duncan, "Class-aware adversarial transformers for medical image segmentation," *Advances in neural information processing systems*, vol. 35, pp. 29 582–29 596, 2022.

[59] M. M. Rahman and R. Marculescu, "Medical image segmentation via cascaded attention decoding," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 6222–6231.

[60] F. Bougourzi, F. Dornaika, A. Taleb-Ahmed, and V. Truong Hoang, "Rethinking attention gated with hybrid dual pyramid transformer-cnn for generalized segmentation in medical imaging," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 243–258.

[61] H. Sun, J. Xu, and Y. Duan, "Paratranscnn: Parallelized transcnn encoder for medical image segmentation," *arXiv preprint arXiv:2401.15307*, 2024.

[62] Y. Gao, S. Zhang, L. Shi, G. Zhao, and Y. Shi, "Collaborative transformer u-shaped network for medical image segmentation," *Applied Soft Computing*, vol. 173, p. 112841, 2025.

[63] S. Yan, B. Yang, A. Chen, X. Zhao, and S. Zhang, "Multi-scale convolutional attention frequency-enhanced transformer network for medical image segmentation," *Information Fusion*, vol. 119, p. 103019, 2025.

[64] C. Huang, Z. Wu, H. Xi, and J. Zhu, "kmaxu: Medical image segmentation u-net with k-means mask transformer and contrastive cluster assignment," *Pattern Recognition*, vol. 161, p. 111274, 2025.