Small Sample Image Segmentation By Coupling Convolutions and Transformers

Hao Qi, Huiyu Zhou, Junyu Dong, *Member, IEEE*, and Xinghui Dong, *Member, IEEE*

***Abstract*—Compared with natural image segmentation, small sample image segmentation tasks, such as medical image segmentation and defect detection, have been less studied. Recent studies made efforts on bringing together Convolutional Neural Networks (CNNs) and Transformers in a serial or interleaved architecture in order to incorporate long-range dependencies into the features extracted using CNNs. In this study, we argue that these architectures limit the capability of the combination of CNNs and Transformers. To this end, we propose a dual-stream small sample image segmentation network, namely, the Interactive Coupling of Convolutions and Transformers Based UNet (ICCT-UNet)[1], motivated by the success achieved using the UNet in the scenario of small sample image segmentation. Within this network, a CNN stream is paralleled with a Transformer stream while maintaining feature exchange inside each block through the proposed Window-Based Multi-head Cross-Attention (W-MHCA) mechanism. To derive an overall segmentation, the features learned by both the streams are further fused using a Residual Fusion Module (RFM). Experimental results show that the ICCT-UNet outperforms, or at least performs comparably to, its counterparts on eight sets of medical and defective images. These promising results should be attributed to the effective combination of the local and global features fulfilled by the proposed interactive coupling method.**

***Index Terms*—Image Segmentation, Convolutional Neural Networks, Transformers, Cross-Attention.**

I. Introduction

SEGMENTATION of natural images has been well studied [1]–[7]. However, this is not the case for image segmentation tasks with a small data set, for example, medical image segmentation and defect detection. Medical images captured by X-Ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound, etc., have been widely used in clinical medicine. In practice, organ or lesion segmentation can assist clinicians to make a more precise diagnosis, design a more appropriate surgical plan and propose treatment strategies. On the other hand, defect detection plays an important role in Non-Destructive Testing (NDT) which is critical to the automatic production process and can significantly reduce the production cost [8].

This study was in part supported by the National Natural Science Foundation of China (NSFC) (No. 42176196) and was in part supported by the Young Taishan Scholars Program (No. tsqn201909060) (Corresponding author: Xinghui Dong).

H. Qi, J. Dong and X. Dong are with the School of Computer Science and Technology, Ocean University of China, Qingdao, 266100. (e-mail: qihao@stu.ouc.edu.cn, dongjunyu@ouc.edu.cn, xinghui.dong@ouc.edu.cn). H. Zhou is with the School of Computing and Mathematical Sciences, University of Leicester, LE1 7RH Leicester, U.K. (e-mail: hz143@leicester.ac.uk).

[1]Code and models are available at https://indtlab.github.io/projects/ICCTUNet.

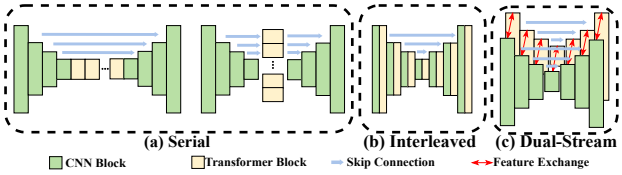

Fig. 1. Comparison of different types of hybrid structures: (a) two serial structures, (b) the interleaved structure and (c) the dual-stream structure.

Although a large number of CNN methods [9]–[15] have been proposed for those tasks, the progress is still slower than that of natural image segmentation. The inherent inductive bias helps these methods learn effective representations from a relatively small data set. However, the limited receptive field of those methods normally prevents them from capturing long-range dependencies [16]. This issue impairs the performance of the CNN methods. On the other hand, Transformer [17], had been introduced into vision tasks and promptly became an alternative to CNNs [18], [19]. Due to the self-attention mechanism, Transformers can be used to extract the context information. However, the strengths of Transformer methods cannot be sufficiently exploited when only a small data set is available [18].

The dilemma can be attributed to two challenges. (**i**) The lack of training images is the Achilles' heel of those methods, which requires that the more powerful priors are encoded in the model while the model is remained appropriately concise in order to avoid over-fitting [20]. (**ii**) Compared with segmentation of natural images, the special image modality and visual content contained in the medical or NDT images result in the low discriminative semantic boundary [21]–[23], which requires the network to preserve a precise representation of the local structure and use a large receptive field to aggregate the context information at the same time.

It has been revealed that the characteristics of CNNs, such as finer local features, shift-invariance and hierarchical representation, can boost the performance of Transformers [24], [25]. Inspired by this finding, existing image segmentation methods, such as TransUNet [26] and MissFormer [27], made efforts on alleviating the two challenges by bringing together CNNs and Transformers to introduce the inductive bias and enhance the ability to model long-range dependencies. However, the serial structure (see Fig. 1(a)) used by these methods limits the exertion of the complementary action of the two sides.

Furthermore, an interleaved [28] (see Fig. 1(b)) or disentangled [29] structure has been used to overcome the limi-

tations of the serial structure. Due to the ignorance of the distinct difference between the features learned by CNNs and Transformers [30], however, those structures may underutilize the potential of these features when they simplistically mix them together. Recently, efforts have also been made in image classification by performing feature communication between CNNs and Transformers [31], [32].

Therefore, we are motivated to propose a novel dual-stream image segmentation network, namely, the Interactive Coupling of Convolutions and Transformers Based UNet (ICCT-UNet) (see Figs. 1(c) or 2(a)), to effectively exploit the locality and globality for image segmentation on small data sets. A new basic block (see Fig. 2(c)) is designed, which comprises a parallel CNN sub-block and a Transformer sub-block while enabling feature exchange between them, to extract local and global features respectively. Within each block, the CNN sub-block is able to receive the global representation from the Transformer sub-block to increase the awareness of the global context, at the same time, the local features can be injected into the Transformer sub-block from the CNN sub-block to help it learn from a small data set. In this case, both the CNN and Transform streams are able to facilitate the other side and a complementary action is achieved.

Instead of fusing the outputs of both the sub-blocks into a single set of feature maps [29], we link the same type of sub-blocks across all the basic blocks, deriving two individual streams. Consequently, the potential of both CNNs and Transformers will be further exploited. We also design a Window-Based Multi-head Cross-Attention (W-MHCA) mechanism (see Fig. 2(d)) for feature exchange. Thanks to the W-MHCA, both the streams can dynamically exchange features at a reasonable computational cost. Two classifiers are placed behind the final decoder block, which introduce the supervision for the two streams and predict two logit maps, respectively. To exploit the features learned by the two streams, we further develop a Residual Fusion Module (RFM). This module fuses these features using residual learning and is appended to the ICCT-UNet. As a result, a third logit map is produced.

Our method is able to utilize the merits of CNNs and Transformers by interactively coupling them. To our knowledge, both the proposed dual-stream network and W-MHCA have not been applied to image segmentation before. Our contributions can be summarized as fourfold.

- We introduce a novel dual-stream image segmentation network, i.e., ICCT-UNet, in which the CNN and Transformer streams exchange features and boost the other side, to effectively exploit both the locality and long-range dependencies. In addition, we design an RFM by leveraging residual learning to integrate the features extracted at the two streams, which normally produces the better prediction than that derived using a single stream.
- To effectively couple both the streams with the relatively low memory and computational demand, we propose a W-MHCA mechanism, which outperforms the addition or concatenation fusion and cross-attention [17] approaches.
- We build a 3D version of the proposed network, which can be applied to volumetric segmentation tasks.
- We demonstrate the effectiveness and generalization of the proposed method on five medical data sets and three defect data sets by experimentation. The results can be used as baselines by the community.

The remainder of this paper is organized as follows. We review the related literature in Section II. In Section III, our methodology is introduced. The experimental setup and results are reported in Sections IV and V respectively while detailed ablation studies are performed in Section VI. Finally, we draw our conclusion in Section VII.

II. Related Work

A. CNN-Based Methods

In [33], a series of Fully Convolutional Networks (FCNs) were evaluated in the scenario of small sample medical image segmentation. As one of the most popular medical image segmentation networks, UNet [9] and its variants [10]–[13], [34]–[37] have been extensively used. The symmetric U-shaped encoder-decoder structure greatly inspired the community. Motivated by the ResNet [38], Xiao et al. [10] introduced residual learning into the UNet by building a U-shaped network on top of residual blocks. In [36], a nested U-shaped network, i.e., UNet++, was proposed, which comprised multiple sub-UNets. Besides, several studies stacked two encoder-decoder networks, e.g., XNet [12] and DoubleU-Net [37], in order to achieve the better performance. In [13], ERDUNet was developed on top of a Context Enhanced Encoder (CEE) module, to extract the global context and leverage the features extracted at different levels for the finer result. Jin et al. [39] designed the DUNet by building a U-shaped network using the deformable convolution for retinal vessel segmentation.

The CNN techniques have also been widely explored in the field of defect detection. In [40], the features extracted using a pre-trained UNet [9] were used together with the random forest classifier for small defect detection. To classify the crack pixels from the background, a full convolutional structure was developed, namely, CrackNet [41]. Zou et al. [15] proposed a deep supervision based feature fusion framework, referred to as DeepCrack, for crack segmentation. Recently, Dong et al. [42] utilized the encoder trained using a synthetic data set for defect detection and classification. In [43], a multi-task framework was also introduced for these tasks, which exploited both the autoencoder and the one-class classifier.

Although the above methods have made great progress, the intrinsic local inductive bias of CNNs restricts the size of receptive fields. As a result, the performance was limited.

B. Transformer-Based Methods

Compared with the CNN-based methods, Transformer [17], which is built upon the self-attention mechanism, naturally owns the ability to capture long-range dependencies. Hence, Transformer-based methods usually showed the better, or at least the comparable, performance in high-level vision tasks [19], [24], [44], [45]. Chen et al. [26] conducted systematic experiments on Vision Transformer (ViT) [18] methods for

medical image segmentation. It was shown that the vanilla ViT can only work with image patches because of the quadratic complexity of the self-attention mechanism. This issue resulted in the absence of the locality and multi-scale features.

To address the issue, Cao et al. [46] proposed a pure Transformer-based UNet, namely, SwinUNet, on top of the Swin Transformer [24] blocks. Since self-attention was constrained within local windows, the SwinUNet reduced the computational cost. Due to the hierarchical representation, the performance was further improved. The Transformer techniques were also used for defect detection. In [47], a Transformer network, i.e., CrackFormer, was developed using the proposed content-based self-attention blocks. To exploit the global context for road surface crack detection, Chen et al. [48] proposed the LECSFormer on top of the window-based self-attention and token rearrangement techniques.

These studies have shown the applicability of Transformers to different tasks. Since Transformers do not contain the strong inductive bias, the superiority to CNNs mainly depends on the large scale of the training data [49]. However, this is normally not the case for either medical image segmentation or defect detection. Although the weights pre-trained using natural images have been used to accelerate convergence [26], [46], the significant domain-shift between these images and the medical or defective images limits the performance.

C. Hybrid Methods

To explore the merits of both CNNs and Transformers, the emphasis was put on the integration of them. It is an intuitive design to place Transformers behind CNNs, to compute global characteristics from the local features learned by CNNs. The TransUnet [26] utilized a ResNet [38] and a Transformer [18] as the encoder and bottleneck respectively. However, this network imprisoned the Transformer at the smallest scale and neglected the multi-scale features learned by CNNs. The following studies [27], [50]–[52] were devoted to solving the weakness of the TransUNet. Among these studies, the MCTrans [51], MissFormer [27] and ScaleFormer [52] adopted a similar design philosophy in which Transformers were used to model the scale-wise relationship, while the UTNet [50] inserted a self-attention module in each block which led to an interleaved hybrid structure. Recently, Guo et al. [29] proposed the UNet-2022 by leveraging a disentangled structure to design the basic block. In [14], a hybrid network, namely, MDAL was proposed by serially integrating convolutional blocks and the self-attention mechanism.

D. 3D Medical Image Segmentation

Some medical image modalities, e.g., MRI images, offer not only the spatial data but also the axial information. 3D or voxel segmentation typically leverages both the sorts of data for the voxel-wise prediction. These methods normally employ 3D convolutional blocks within a U-shaped encoder-decoder structure [53]–[56]. Despite the improved performance has been achieved using those methods, the locality limitation still remains in them due to the multi-dimensional aggregation scope of the 3D convolution.

To address this issue, many studies introduced self-attention mechanisms to 3D medical image segmentation [28], [57]–[59]. In [58], Hatamizadeh et al. proposed UNETR, which combined a Transformer encoder and a CNN decoder. Zhou et al. [28] designed a hybrid encoder-decoder architecture, referred to as nnFormer, by stacking convolutional and Transformer blocks in an interleaved manner. It was shown that the joint use of the locality and long-range dependencies was beneficial for 3D medical image segmentation.

Guo et al. [29] disentangled the CNN and Transformer sub-blocks in order to balance the use of the locality and globality. However, the unidirectional fusion mechanism lacked adaptability. Pioneering studies [30], [49] have shown distinct feature differences between CNNs and Transformers. In this context, an even fusion may hinder the full utilization of them [60]. Recently, bi-directional CNN-Transformer communication has been applied to image classification [31], [32]. These studies mainly paid attention to the introduction of the global context, while ignoring the utilization of hierarchical representations which are essential to image segmentation. In contrast, we propose a dual-stream encoder-decoder network (see Fig. 1(c)) by coupling CNNs and Transformers while enabling them to exchange features using a new Window-Based Multi-head Cross-Attention (W-MHCA) mechanism.

III. Methodology

The architecture of the proposed method is shown in Fig. 2(a), which comprises an Interactive Coupling of Convolutions and Transformers Based UNet (ICCT-UNet) and a Residual Fusion Module (RFM). As can be seen, the ICCT-UNet contains two parallel streams: CNNs and Transformers. In terms of these streams, two sub-blocks in the same block can exchange features. Each stream produces a single logit map. To fuse the features learned by those streams, the RFM is appended to the ICCT-UNet. A third logit map is predicted by it. Regarding the two streams and RFM, three loss terms are computed respectively. The entire network can be end-to-end trained by summing up these terms. However, we found that they were difficult to weight. To further explore the capabilities of the two streams and RFM, they can be separately trained by applying the stop-gradient technique to the beginning of the RFM. In addition, we implement a 3D variant of the proposed network, which can be used for 3D image segmentation.

A. ICCT-UNet

Within the U-shaped ICCT-UNet, the encoder contains a series of blocks: Enc_i ($i \in \{0, 1, 2, 3, 4\}$) while the decoder consists of a different set of blocks: Dec_j ($j \in \{3, 2, 1, 0\}$). The CNN stream is formulated with the sub-blocks which are similar to those used by the UNet [9]. On the other hand, the Transformer stream is built on top of the sub-blocks that SwinUNet [46] used. With regard to both the streams, the features computed using the two sub-blocks in the same block are exchanged through the Window-Based Cross-Attention (W-MHCA) mechanism. An individual logit map is predicted by each stream.

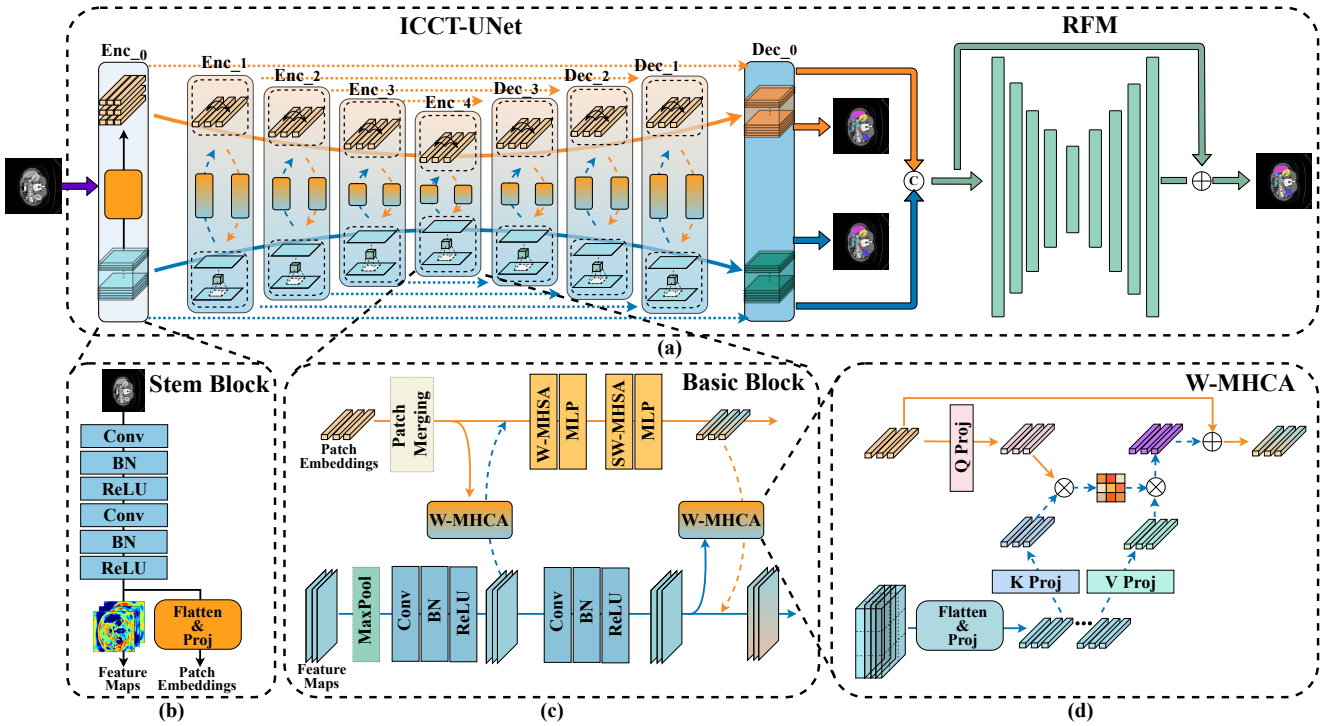

Fig. 2. Illustration of the proposed image segmentation method which contains an ICCT-UNet and a Residual Fusion Module (RFM). (a) shows the network architecture, while (b) and (c) present the internal structure of the stem block and the basic block respectively, and (d) displays the pipeline of W-MHCA.

Encoder. To derive an initial representation for each stream, we use a stem block (see Fig. 2(b)) denoted as Enc_0, which contains two consecutive convolutional layers and each layer is followed by a Batch Normalization (BN) layer and a ReLU activate function. As a result, the input image is transformed to a set of feature maps which have the same resolution as the image. It should be noted that this processing is different from that performed using the first block of the TransUNet [26] and SwinUNet [46], which down-samples the input. The feature maps are directly fed into the CNN stream. Meanwhile, they are flattened and processed using a linear projection. As a result, a set of initial patch-embeddings are generated. They are then sent to the Transformer stream.

Following the Enc_0, there are three additional blocks and a bottleneck, Enc_i ($i \in \{1, 2, 3, 4\}$), in the encoder, to learn representations at different semantic levels. In each block (see Fig. 2(c)), the CNN sub-block contains two units ($unit(\cdot)$), each of which comprises a set of $Conv$-BN-$ReLU$ operations; while the Transformer sub-block includes a Shifted Window-Based Multi-head Self-Attention (SW-MHSA) unit and a two-layer Multilayer Perceptron (MLP). (The W-MHSA is a zero-shift SW-MHSA). Given that $X^{i-1}_{CNN} \in \mathbb{R}^{C\times H\times W}$ denotes the feature maps extracted by the CNN sub-block in the Enc_{i-1} and $\hat{X^{i}_{CNN}}$ represents the feature maps produced by the *MaxPooling* operation, the computation of the CNN sub-block in the Enc_i can be expressed as:

$$\hat{X^{i}_{CNN}} = MaxPooling(X^{i-1}_{CNN}), \tag{1}$$

$$X^{i}_{CNN} = unit(unit(\hat{X^{i}_{CNN}})), \tag{2}$$

where $X^{i}_{CNN} \in \mathbb{R}^{2C\times \frac{H}{2}\times \frac{W}{2}}$.

Let $Y^{i-1}_{Trans} \in \mathbb{R}^{HW\times C}$ denote the patch-embeddings produced by the Transformer sub-block in the Enc_{i-1}, $\hat{Y}^{i}_{Trans}$ stand for the downsampled patch-embeddings and $\tilde{Y}^{i}_{Trans}$ represent the intermediate result, the computation of the sub-block in the Enc_i can be expressed as:

$$\hat{Y}^{i}_{Trans} = PatchMerging(Y^{i-1}_{Trans}), \tag{3}$$

$$\tilde{Y}^{i}_{Trans} = W\text{-}MHSA(LN(\hat{Y}^{i}_{Trans})) + \hat{Y}^{i}_{Trans}, \tag{4}$$

$$\tilde{Y}^{i}_{Trans} = MLP(LN(\tilde{Y}^{i}_{Trans})) + \tilde{Y}^{i}_{Trans}, \tag{5}$$

$$\tilde{Y}^{i}_{Trans} = SW\text{-}MHSA(LN(\tilde{Y}^{i}_{Trans})) + \tilde{Y}^{i}_{Trans}, \tag{6}$$

$$Y^{i}_{Trans} = MLP(LN(\tilde{Y}^{i}_{Trans})) + \tilde{Y}^{i}_{Trans}, \tag{7}$$

where $Y^{i}_{Trans} \in \mathbb{R}^{\frac{HW}{4}\times 2C}$ and $LN(\cdot)$ denotes Layer Normalization.

In the bottleneck, i.e., Enc_4, the computation process is nearly the same as the previous blocks except that the number of channels of the feature maps outputted is equal to that of the input feature maps in the CNN sub-block.

Decoder. The decoder of the ICCT-UNet contains four blocks: Dec_j ($j \in \{3, 2, 1, 0\}$). In the Dec_0, both the CNN and Transformer sub-blocks are the same as the CNN sub-block contained in an encoder block. In terms of each stream, however, the sub-blocks in the Dec_3 to Dec_1 are the same as those comprised in the Enc_3 to Enc_1 respectively. Regarding each block, the input of the CNN sub-block or the Transformer sub-block is the concatenation of two sets of feature maps. One set is the result obtained by applying bilinear

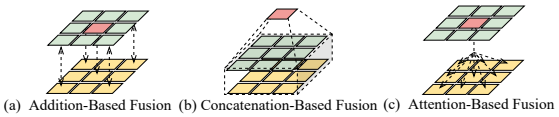

Fig. 3. Different approaches for feature fusion.

interpolation to the feature maps produced by the previous sub-block in the same stream, while the other set is generated by the sub-block at the same level in the corresponding stream of the encoder. In particular, a linear projection is used in the Transformer sub-blocks to adjust the dimension of the input. A 1×1 convolutional layer is applied to the output of each sub-block in the Dec_0 and the result is a logit map. To further reduce the possibility of over-fitting, we can apply the Spatial Dropout [61] technique to the ICCT-UNet. For each stream, we use a Spatial Dropout unit in the Dec_j ($j \in \{3, 2, 1, 0\}$).

As shown in Fig. 2(a), the bi-directional interactions between the two streams take place in each block except the Enc_0 and Dec_0. Due to the local inductive bias of CNNs, the training of them is easier than that of Transformers. Thus, the injection $CNNs \rightarrow Trans$ can accelerate the training of the Transformer stream. On the other hand, Transformer sub-blocks are able to incorporate the global information into the CNN sub-blocks through the $Trans \rightarrow CNNs$ pathway, which is useful for the CNN sub-blocks to overcome their limited receptive fields. Hence, the ICCT-UNet can overwhelm the data scale limitation and learn from scratch with a small data set. Since it integrates both local characteristics and the global context, the more accurate segmentation can be derived.

W-MHCA. In essence, the feature exchange between the two streams can be treated as feature fusion. Given two sets of features, the commonly-used fusion methods include the point-wise addition and the concatenation followed by a weighted summing (e.g., convolution), as shown in Figs. 3(a) and (b) respectively. Compared with the addition method, the concatenation method can aggregate more features using a large neighbourhood but it requires more parameters.

In [31], a Feature Coupling Unit (FCU) was used to fill the semantic gap between the CNN and Transformer streams for image classification, in which feature fusion was fulfilled by addition. More sophisticatedly, a simplified version of cross-attention [17] was also utilized in order to fuse the features extracted using CNNs and Transformers [32]. As shown in Fig. 3(c), the stream, which intakes the information from the other, uses its own features as the query and utilizes the features of the other as the key and value. Compared with the addition and concatenation methods, cross-attention is able to dynamically fuse features because fusion weights are computed according to the input. Since the projections in the cross-attention unit are shared by all tokens, less parameters are used to aggregate a large number of features than those that the concatenation method needs.

The cross-attention computation was conducted across all the tokens at each resolution level [32]. Although this global computation works well with image classification which only needs the image-level representation, it is unsuitable for dense prediction tasks because they use fine-grained representations to precisely localize objects. Considering feature exchange is conducted at each level, the global computation will require a huge amount of computational resource when it is computed at the high resolution level. In addition, the global cross-attention mechanism may produce redundant features because the irrelevant information is introduced. To address these issues, we propose a new Window-Based Multi-head Cross-Attention (W-MHCA) mechanism on top of an improved attention method [62]. This mechanism is able to efficiently bridge the CNN and Transformer streams (see Fig. 2(d)).

Given the features extracted in the CNN and Transformer sub-blocks of the same block, we split these into $w \times w$ non-overlapping windows. Then the multi-head cross-attention is computed within these windows. Compared with the global cross-attention mechanism [32], the W-MHCA has a reasonable computational complexity while reducing the redundant computation on the irrelevant tokens.

When feature fusion is performed from the CNN stream to the Transformer stream, for example, the features produced in the CNN stream are used as the key and value while the features extracted in the Transformer stream are used as the query. Similar to the MHSA [18], the W-MHCA contains multiple attention heads. The results produced by these heads are fused using an output projection W_O. Let X_{CNN} and $Y_{Trans} \in \mathbb{R}^{w^2 \times d}$ be the d-dimensional tokens within a window in the CNN and Transformer streams respectively, and W_Q^j, W_K^j and W_V^j be the query, key and value projections of the j-th attention head, the output of this head, i.e., $\hat{Y}_{Trans}^j$, can be computed as:

$$\hat{Y}_{Trans}^j = Softmax(\frac{QK^T}{s})V, \tag{8}$$

$$Q = W_Q^j Y_{Trans}, K = W_K^j X_{CNN}, V = W_V^j X_{CNN}, \tag{9}$$

where s is the temperature term of the softmax function and is defined as a learnable parameter in order to avoid the attention deterioration problem [62]. Finally, the output of the W-MHCA with h attention heads is computed as:

$$Y_{Trans}^{Fuse} = W_O \hat{Y}_{Trans}^{Fuse} + Y_{Trans}^j, \tag{10}$$

$$\hat{Y}_{Trans}^{Fuse} = concatenate(\hat{Y}_{Trans}^1, \hat{X}_{Trans}^2, ..., \hat{X}_{Trans}^h). \tag{11}$$

B. Residual Fusion Module

The two streams of the ICCT-UNet produce two different logit maps respectively. To jointly exploit the discriminant ability of these streams, we design a simple sub-network, referred to as the Residual Fuse Module (RFM), which performs residual learning on the concatenation of the features extracted at the end of the two streams. The RFM aims to fuse these features and generate the finer logit map.

Motivated by the ResNet [38], the computation of RFM can be expressed as:

$$X_{out} = X_{in} + \mathcal{F}(X_{in}), \tag{12}$$

where X_{in} is the concatenation of the features extracted in the two streams of the decoder and $\mathcal{F}(\cdot)$ stands for the RFM.

The RFM is also a U-shaped encoder-decoder network with skip connections. Each block contains a convolutional layer, a batch normalization layer and a ReLU activate function. Two additional 3×3 convolutional layers are placed in front of the first block of the encoder and behind the last block of the decoder respectively. The encoder comprises five blocks in which the last block is used as the bottleneck. With the spatial resolution of feature maps decreases, the number of channels is kept constant, i.e., 64. Similarly, the decoder consists of four blocks. Finally, we use a 1×1 convolutional layer to perform the pixel-wise classification. The result is a third logit map in addition to the two logit maps produced by the ICCT-UNet.

However, it is challenging to obtain the optimal weights in order to balance the end-to-end training processes of the ICCT-UNet and RFM. In particular, inappropriate weights may disturb the training of the ICCT-UNet. As a result, imperfect representations are learned by the two streams. Since these representations are fed into the RFM, the performance of it will be impaired. Although this issue can be alleviated by setting a small weight for the loss of the RFM, it may result in the poor training of the RFM. To get rid of the dilemma, we apply the stop-gradient operation to the beginning of the RFM. In this case, the ICCT-UNet and RFM can be trained independently while they will be optimized appropriately.

C. 3D Variant

The aforementioned ICCT-UNet and RFM can be easily modified for 3D image segmentation tasks. We first replace the convolutional layers and batch normalization with the corresponding 3D versions. In this case, both the CNN stream and RFM are reformulated. To reformulate the Transformer stream and W-MHCA, we then use local volumes to replace the local windows required for computing the self-attention and cross-attention mechanisms. As a result, a 3D variant of the proposed network is built.

IV. Experimental Setup

We performed a series of experiments on eight data sets. In this section, we will briefly introduce the data sets, experimental setup and implementation details.

A. Data Sets

For 2D medical image segmentation, our method was tested with four data sets, including *Synapse* [63], *ACDC* [22], *ISIC 2018* [64] and *BUSI* [65]. The *Synapse* data set contains 30 abdominal CT scans in total. Both the Dice Score (DSC) and 95% Hausdorff Distance (HD95) were computed on eight main organs. The *ACDC* data set comprises 100 cases for segmentation of the Left Ventricle (LV), Right Ventricle (RV) and Myocardium (MYO). Only the DSC was calculated for this data set by following the previous work [26], [28], [46]. In total, 2,594 dermatologic images of the skin lesion are included in the *ISIC 2018* data set. Regarding the *BUSI* data set, 647 benign and malignant cancer ultrasound images were used. Both IoU and the F1 score were used as the performance measures for the *ISIC* and *BUSI* data sets.

With regard to 3D medical image segmentation, we used three data sets, including *Synapse* [63], *ACDC* [22] and *MSD* [66]. The *MSD* data set contains 484 MRI images for brain tumor segmentation. We examined the segmentation results of the Whole Tumor (WT), Enhancing Tumor (ET) and Tumor Core (TC) in terms of the DSC and HD95 metrics.

For defect detection experiments, we evaluated our method on three data sets: *CFD* [67], *MT* [68] and *KSDD* [69]. The *CFD* data set includes 183 road surface crack images with the annotated cracks. Five sets of magnetic tile surface defect images and one set of defect-free images are comprised of the *MT* data set. We used 172 images of the blowhole and crack defects by following the original setup [68]. The *KSDD* data set contains 52 annotated defect images. We computed both the IoU and AUC values for the three data sets.

Regarding the *Synapse* [63], *ACDC* [22] and *MSD* [66] data sets, the training/testing split was kept the same as that used in the previous studies [26], [28], [46]. For the *ISIC 2018* [64], *BUSI* [65] and three defect data sets, we split these into five folds and conducted cross-validation experiments. The average was computed across the results obtained in the five folds. For more details, please refer to Table III.

B. Experimental Setup

For the 2D image segmentation experiments, we resized the images to the resolution of 224×224 pixels. Only the cross-entropy loss was used to train both the ICCT-UNet and RFM. The AdamW [70] optimizer was utilized to optimize the network. We also employed the poly learning rate attenuation strategy [26], which is expressed as:

$$lr = initial\_lr \times (1 - \frac{iters}{iters\_total})^{0.9}, \tag{13}$$

where $iters$ means the number of iterations that have been completed and $iters\_total$ denotes the number of iterations required in the experiment. Table I shows the details of the experimental setup used for each data set.

Regarding the 3D image segmentation experiments, we used the same training tactics as those used by the nnFormer [28], including the spacing and cropping methods, combined loss of the cross-entropy and dice losses, SGD optimizer and Deep Supervision technique. Please refer to Table II for more details.

To alleviate the potential overfitting problem, we used the same data augmentation operations as those utilized in [26], [46], including horizontal flip, vertical flip and random rotation, for the 2D image segmentation task, while we used the same data augmentation operations as those employed by the nnFormer [28] in the 3D image segmentation task. We also used Spatial Dropout [61] as an extra regularization method for training our network.

C. Implementation Details

In the remainder of this paper, we use ICCT-UNet-X to denote the proposed ICCT-UNet with a specific number of channels of the feature maps extracted using the Enc_0. The predictions of the CNN stream, the Transformer stream and the RFM are denoted as ICCT-UNet-X-C, ICCT-UNet-X-T and

TABLE I
THE SETUP UTILIZED FOR THE SEVEN DATA SETS USED IN 2D IMAGE SEGMENTATION, INCLUDING THE BATCH SIZE, MAXIMAL EPOCHS, INITIAL LEARNING RATE, WEIGHT DECAY AND SPATIAL DROPOUT RATE.

	Synapse	ACDC	BUSI	ISIC	MT	CFD	KSDD
Batch Size	4	4	4	4	4	4	4
Max. Epochs	450	450	400	400	200	200	200
Learning Rate	5e-4	5e-4	5e-4	5e-4	5e-4	5e-4	5e-4
Weight Decay	5e-2	5e-2	1e-2	1e-2	1e-2	3e-1	3e-1
Spatial Dropout	[0.4,0.4,0.2,0.2]	[0.4,0.4,0.2,0.2]	None	None	None	None	None

TABLE II
THE SETUP USED FOR THE THREE DATA SETS UTILIZED IN 3D MEDICAL IMAGE SEGMENTATION, INCLUDING THE SPACING, MEDIAN SHAPE, CROPPING SIZE, BATCH SIZE, MAXIMAL EPOCHS, INITIAL LEARNING RATE, WEIGHT DECAY, MOMENTUM AND SPATIAL DROPOUT RATE.

	Synpse	ACDC	MSD
Spacing	[0.76, 0.76, 3]	[1.52, 1.52, 6.35]	[1.0, 1.0, 1.0]
Median Shape	512 × 512 × 148	246 × 213 × 13	138 × 170 × 138
Cropping Size	128 × 128 × 64	160 × 160 × 14	128 × 128 × 128
Batch Size	2	4	2
Max. Epochs	1000	1000	1000
Learning Rate	1e-2	1e-2	1e-2
Weight Decay	3e-5	3e-5	3e-5
Momentum	0.99	0.99	0.99
Spatial Dropout	None	None	None

ICCT-UNet-X-F, respectively. Considering the varying scale of different data sets, we used the ICCT-UNet-64 for the *Synapse* [63] and *ACDC* [22] data sets and utilized the ICCT-UNet-32 for the other data sets in the 2D segmentation experiments. In each encoder or decoder block, the numbers of the attention heads in the SW-MHSA and W-MHCA were set to the same value. For the Enc_{i} ($i \in \{1, 2, 3, 4\}$), we used 4, 8, 16 and 32 attention heads in turn. In terms of the Dec_{j} ($j \in \{1, 2, 3\}$), 4, 8, and 16 attention heads were utilized in turn. The size of windows w was set to 7 for the SW-MHSA unit in the Transformer stream and all the W-MHCA units.

Regarding the 3D segmentation experiments, we replaced all the blocks and modules by the corresponding 3D versions. To reduce the computation and memory demand, the stride of the convolutions in the Enc_{0} was set to 2, which generate the smaller feature maps and fewer patch embeddings. Since we followed the experimental setup utilized by the nnFormer [28], the same hyperparameters were used for our Transformer stream to avoid tuning the network.

It should be noted that our Transformer stream was different from the nnFormer [28]. To be specific, we used the Patch-Merging [24] operation to connect different encoder blocks and the locality information was introduced by the CNN stream through the W-MHCA unit. In contrast, the nnFormer used the overlapped convolution. Furthermore, the nnFormer used a skip-attention module [28] to fuse the features between the encoder and decoder while we used the more simplistic skip connection method [9]. In addition, the nnFormer used the transpose convolution in the decoder to upsample feature maps while we employed the naive trilinear interpolation.

All the image segmentation approaches tested in this study were implemented using Python 3.9.7, Pytorch 1.10.2 and Torchvision 0.11.3. The experiments were performed on an NVIDIA Geforce RTX 3090 graphics card.

TABLE III
THE IMAGING TECHNIQUE AND TRAINING/TESTING SPLIT USED FOR EACH OF EIGHT DATA SETS.

	Synapse	ACDC	MSD	BUSI	ISIC	MT	CFD	KSDD
Imaging	CT	MRI	MRI	Ultrasound	Dermoscopy	Camera	Camera	Camera
#Training	18	70	388	517	2075	146	137	41
#Testing	12	20	242	130	519	37	46	11

V. EXPERIMENTAL RESULTS

In this section, we will report the results obtained using five medical and three defect data sets.

A. Medical Image Segmentation

The results obtained using different methods on the *Synapse* [63], *ACDC* [22], *MSD* [66], *BUSI* [65] and *ISIC* [64] data sets are reported in Tables IV, V, VI and VII. As can be seen, the three predictions of our method were superior to, or at least comparable to, different baselines, including CNN methods (e.g., UNet [9] and XNet [12]), Transformer methods (e.g., SwinUNet [46]) and serial hybrid methods (e.g., TransUNet [26], ScaleFormer [52] and MISSFormer [27]) across the five data sets. Our 3D model outperformed its state-of-the-art counterpart, i.e., nnUNet [56], on three data sets. This model also outperformed the interleaved network, nnFormer [28], on the *Synapse* and *MSD* data sets and produced the comparable result on the *ACDC* data set. Compared with the disentangled model, UNet-2022 [29], which was trained using a different setup, our method achieved the comparable performance. Given that our training setup was used, however, our method was superior to the UNet-2022. In addition, the fusion of the two streams normally generated the better result than that produced by each individual stream.

For the *Synapse* data set (see Table IV), our 3D model generated the best result. In particular, the Transformer stream outperformed the state-of-the-art method, nnFormer [28]. Our 2D model produced the second best result. Compared with the results of the UNet-2022 reported in [29], our method was superior on the *Gallbladder*, *Spleen* and *Stomach* images while was comparable on the *Kidney (Right)* and *Liver* images.

The proposed method also achieved comparable results to those produced by its counterparts on the *ACDC* data set (see Table V). It is noteworthy that both the TransUNet [26] and SwinUNet [46] used pre-trained weights as the initialization even if they produced the better results than our method on the *LV* images. In contrast, our method can be trained from scratch on the target data set. Besides, the average performance of the Transformer stream of our method was better than that of the SwinUNet [46] which had a similar architecture.

On the *MSD* data set (see Tabe VI), our method outperformed the state-of-the-art method, nnFormer [28]. Although this method has set a significantly high benchmark, we did not aim to design the most powerful volumetric segmentation approach. Nevertheless, our 3D model, derived by replacing

TABLE IV
COMPARISON OF DIFFERENT METHODS ON THE *Synapse* [63] DATA SET. HERE, * IMPLIES THE RESULTS DERIVED USING A 3D SEGMENTATION MODEL AND † SUGGESTS THE RESULTS OBTAINED USING OUR TRAINING/INFERENCE SETUP. FOR EACH COLUMN, THE BEST, SECOND BEST AND THIRD BEST PERFORMANCES ARE HIGHLIGHTED IN **RED**, *CYAN* AND BLUE FONTS RESPECTIVELY. THIS CONTINUES IN TABLES V, VI, VII AND VIII.

Method	Average		Aotra	Gallbladder	Kidney (Left)	Kidney (Right)	Liver	Pancreas	Spleen	Stomach
	HD95↓	DSC↑								
UNet [9]	39.70 [46]	76.85 [46]	89.07 [46]	69.72 [46]	77.77 [46]	68.60 [46]	93.43 [46]	53.98 [46]	86.67 [46]	75.58 [46]
UNet† [9]	14.84	79.71	89.16	66.52	82.14	70.81	95.10	65.82	87.82	80.35
XNet [12]	11.70	82.70	89.70	65.73	83.90	81.58	95.37	68.27	91.56	85.51
ViT [18] + CUP [26]	36.11 [26]	67.86 [26]	70.19 [26]	45.10 [26]	74.70 [26]	67.40 [26]	91.32 [26]	42.00 [26]	81.75 [26]	70.44 [26]
R50-ViT [18] + CUP [26]	32.87 [26]	71.29 [26]	73.73 [26]	55.13 [26]	75.80 [26]	72.20 [26]	91.51 [26]	45.99 [26]	81.99 [26]	73.95 [26]
TransUNet [26]	31.69 [26]	77.48 [26]	87.23 [26]	63.16 [26]	81.87 [26]	77.02 [26]	94.08 [26]	55.86 [26]	85.08 [26]	75.62 [26]
TransUNet† [26]	17.22	77.39	86.75	70.17	76.91	72.12	93.68	57.93	86.08	75.50
SwinUNet [46]	21.55 [46]	79.13 [46]	85.47 [46]	66.53 [46]	83.28 [46]	79.61 [46]	94.29 [46]	56.58 [46]	90.66 [46]	76.60 [46]
SwinUNet† [46]	18.77	74.12	81.13	60.76	76.07	73.18	93.36	47.81	86.56	74.12
MISSFormer [27]	18.20 [27]	81.96 [27]	86.99 [27]	68.65 [27]	85.21 [27]	82.00 [27]	94.41 [27]	65.67 [27]	91.92 [27]	80.81 [27]
MISSFormer† [27]	18.50	78.55	85.32	65.75	80.79	72.36	95.38	59.05	88.26	81.46
ScaleFormer [52]	16.81 [52]	82.86 [52]	88.73 [52]	**74.97** [52]	86.36 [52]	83.31 [52]	95.12 [52]	64.85 [52]	89.40 [52]	80.14 [52]
ScaleFormer† [52]	17.37	79.03	86.20	60.91	75.81	76.45	95.11	65.04	91.59	81.17
UNet-2022 [29]	16.70 [29]	84.98 [29]	92.10 [29]	69.63 [29]	**88.40** [29]	83.93 [29]	96.02 [29]	75.50 [29]	90.40 [29]	83.86 [29]
UNet-2022† [29]	11.99	80.35	87.66	68.97	82.93	79.79	95.11	59.99	89.29	79.02
UNETR* [58]	22.97 [28]	79.56 [28]	89.99 [28]	60.56 [28]	85.66 [28]	84.80 [28]	94.46 [28]	59.25 [28]	87.81 [28]	73.99 [28]
Swin-UNETR* [59]	15.25	82.43	91.01	65.88	85.68	85.50	95.93	71.45	89.76	74.22
nnUNet* [56]	10.78 [28]	86.99 [28]	*93.01* [28]	71.77 [28]	85.57 [28]	**88.18** [28]	**97.23** [28]	83.01 [28]	91.86 [28]	85.26 [28]
nnFormer* [28]	10.63 [28]	86.57 [28]	92.04 [28]	70.17 [28]	86.57 [28]	86.25 [28]	96.84 [28]	83.35 [28]	90.51 [28]	86.83 [28]
Ours ICCT-UNet-64-C	11.11	84.18	90.64	69.78	86.26	83.81	95.54	72.34	91.80	83.25
Ours ICCT-UNet-64-T	10.47	83.50	90.49	68.74	85.18	83.73	95.46	70.63	91.29	82.47
Ours ICCT-UNet-64-F	11.01	84.60	91.13	71.50	86.58	83.86	95.64	72.20	91.81	84.07
Ours ICCT-UNet-192-C*	8.94	*87.45*	*93.01*	71.43	87.53	87.09	*96.98*	*83.42*	*92.53*	**87.59**
Ours ICCT-UNet-192-T*	*8.91*	87.24	92.86	71.23	87.40	87.03	96.94	82.74	92.51	87.25
Ours ICCT-UNet-192-F*	**8.73**	**87.61**	**93.03**	*72.21*	*87.71*	*87.17*	96.97	**83.65**	**92.61**	*87.54*

TABLE V
COMPARISON OF DIFFERENT METHODS ON THE *ACDC* [22] DATA SET.

Method	Average DSC↑	RV	Myo	LV
UNet [9]	90.98	89.36	88.42	95.17
XNet [12]	91.21	89.62	88.56	95.46
ViT [18] + CUP [26]	81.45 [26]	81.46 [26]	70.71 [26]	92.18 [26]
R50-ViT [18] + CUP [26]	87.57 [26]	86.07 [26]	81.88 [26]	94.75 [26]
SwinUNet [46]	90.00 [46]	88.55 [46]	85.62 [46]	95.83 [46]
SwinUNet† [46]	90.08	89.19	86.86	94.19
TransUNet [26]	89.71 [26]	88.86 [26]	84.53 [26]	95.73 [26]
TransUNet† [26]	90.51	88.96	87.39	95.18
MissFormer [27]	90.86 [27]	89.55 [27]	88.04 [27]	94.99 [27]
MissFormer† [27]	90.86	89.46	87.88	95.24
ScaleFormer [52]	90.17 [52]	87.33 [52]	88.16 [52]	95.04 [52]
ScaleFormer† [52]	86.44	84.02	82.59	92.71
UNet-2022 [29]	**92.83** [29]	**91.04** [29]	**90.97** [29]	**96.49** [29]
UNet-2022† [29]	90.67	89.37	88.06	94.58
UNETR* [58]	88.61 [28]	85.29 [28]	86.52 [28]	94.02 [28]
nnUNet* [56]	91.61 [28]	90.24 [28]	89.24 [28]	95.36 [28]
nnFormer* [28]	*92.01* [28]	*90.94* [28]	*89.58* [28]	*95.65* [28]
Ours ICCT-UNet-64-C	91.61	90.59	88.92	95.31
Ours ICCT-UNet-64-T	91.13	89.89	88.34	95.08
Ours ICCT-UNet-64-F	91.64	90.66	88.94	95.30
Ours ICCT-UNet-96-C*	91.88	90.74	89.33	95.57
Ours ICCT-UNet-96-T*	91.84	90.66	89.27	95.59
Ours ICCT-UNet-96-F*	91.91	90.79	89.34	95.58

TABLE VI
COMPARISON OF DIFFERENT METHODS ON THE *MSD* [66] DATA SET. THE RESULTS SHOWN IN ROWS 1-7 ARE DERIVED FROM [28].

Method	Average		WT		ET		TC	
	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑
SETR NUP [45]	13.78	63.7	14.42	69.7	11.72	54.4	15.19	66.9
SETR PUP [45]	14.01	63.8	15.25	69.6	11.76	54.9	15.02	67.0
SETR MLA [45]	13.49	63.9	15.50	69.8	10.24	55.4	14.72	66.5
TransBTS* [26]	9.65	69.6	10.03	77.9	9.97	57.4	8.95	73.5
UNETR* [58]	8.82	71.1	8.27	78.9	9.35	58.5	8.85	76.1
nnUNet* [56]	4.20	86.1	3.64	**92.0**	4.06	81.0	4.91	85.4
nnFormer* [28]	4.05	86.4	3.80	91.3	**3.87**	81.8	4.49	**86.0**
Swin-UNETR* [59]	4.07	85.5	3.71	91.4	3.57	79.2	4.91	85.8
UNet [9]	6.18	71.4	6.73	80.1	5.03	59.2	6.77	74.9
SwinUNet [46]	8.54	65.4	12.78	73.0	6.55	51.2	6.29	72.2
TransUNet [26]	6.41	70.4	6.90	78.7	5.17	56.6	7.16	75.9
Ours ICCT-UNet-96-C*	**3.97**	86.3	*3.59*	91.8	*4.30*	81.3	**3.59**	85.6
Ours ICCT-UNet-96-T*	*4.02*	*86.5*	**3.37**	*91.9*	4.31	*81.9*	*4.39*	85.7
Ours ICCT-UNet-96-F*	4.22	**86.6**	3.63	**92.0**	4.40	**82.0**	4.62	*85.8*

the units by their 3D counterparts without making more modifications, still achieved the comparable results. The effectiveness and generalization of our method has been shown.

When the *BUSI* and *ISIC* data sets (see Table VII) were used, our method always outperformed the baselines no matter IoU or the F1 Score was considered. Particularly, each prediction of our method was the best. In contrast, the SwinUNet [46] struggled with segmenting the *BUSI* images.

It is noteworthy that the UNet-2022 used more data augmen-

TABLE VII
COMPARISON OF DIFFERENT METHODS ON THE *BUSI* [65] AND *ISIC 2018* [64] DATA SETS.

Method	BUSI [65]		ISIC 2018 [64]	
	IoU↑	F1↑	IoU↑	F1↑
UNet [9]	68.16	77.46	81.21	89.18
XNet [12]	66.87	77.14	81.07	88.99
SwinUNet [46]	33.97	47.77	79.24	88.32
TransUNet [26]	66.67	76.80	81.49	89.83
MissFormer [27]	58.22	68.83	80.54	89.69
UNet-2022 [29]	63.52	75.41	81.30	89.62
Ours ICCT-UNet-32-C	*69.89*	*79.49*	*81.93*	*90.10*
Ours ICCT-UNet-32-T	69.53	79.32	81.87	90.08
Ours ICCT-UNet-32-F	**70.27**	**79.70**	**82.11**	**90.23**

TABLE VIII
COMPARISON OF DIFFERENT METHODS ON THE *CFD* [67], *MT* [68] AND *KSDD* [69] DATA SETS.

Method	CFD [67]		MT [68]		KSDD [69]	
	IoU↑	AUC↑	IoU↑	AUC↑	IoU↑	AUC↑
UNet [9]	48.14	0.974	56.83	*0.998*	*58.74*	*0.996*
XNet [12]	48.40	*0.985*	53.77	0.997	57.85	0.992
SwinUNet [46]	31.18	0.952	43.93	0.991	30.47	0.964
TransUNet [26]	47.21	0.984	*60.35*	**0.999**	44.02	0.928
MissFormer [27]	45.18	0.981	37.73	0.894	35.72	0.892
UNet-2022 [29]	48.02	*0.985*	**65.52**	**0.999**	54.67	**0.997**
Ours ICCT-UNet-32-C	47.03	0.983	57.69	0.997	56.72	0.990
Ours ICCT-UNet-32-T	*48.45*	0.981	55.47	0.997	55.26	0.991
Ours ICCT-UNet-32-F	**49.82**	**0.987**	58.36	**0.999**	**59.40**	**0.997**

tation operations and the more complicated training/inference setup on the *Synapse* and *ACDC* data sets, including a multi-term loss function, a deep supervision method and a patch-fusion-based inference strategy. The performance gain of the UNet-2022 may be attributed to these techniques. In contrast, our training setup is much simpler and our model can perform inference over an image rather than a set of image patches. When our training/inference setup was used with the UNet-2022, a significant performance degradation was observed.

B. Defect Detection

Defect detection was also performed based on the 2D image segmentation task. Regarding the three defect data sets: *CFD* [67], *MT* [68] and *KSDD* [69], the results obtained using different methods are shown in Table VIII. Again, the results produced by our method were better than, or at least comparable to, those derived using the baselines.

C. Visualization of Segmentation

In Fig. 4, we visualize the results obtained using six networks on an image of eight data sets. It can be observed that our method was able to localize the organs and defects in various scales with the higher accuracy than its counterparts. In particular, UNet [9] struggled to handle complex organ boundaries and tended to produce over-segmentation results. This observation should be attributed to the limited effective receptive field which hinders UNet from capturing the sufficient semantic context. Although the pure Transformer-based

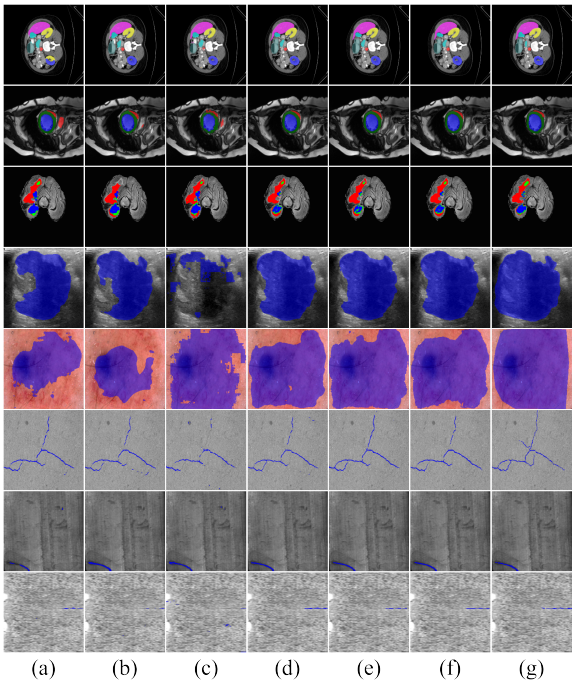

Fig. 4. The results obtained on eight data sets, including *Synapse* [63], *ACDC* [22], *MSD* [66], *BUSI* [65], *ISIC* [64], *CFD* [67], *MT* [68] and *KSDD* [69], are shown in the above eight rows respectively (For the best viewing, please zoom in.). Each row displays the results derived using (a) UNet [9], (b) TransUNet [26], (c) SwinUNet [46], (d) ICCT-UNet-T, (e) ICCT-UNet-C, (f) ICCT-UNet-F and (g) the ground-truth in turn.

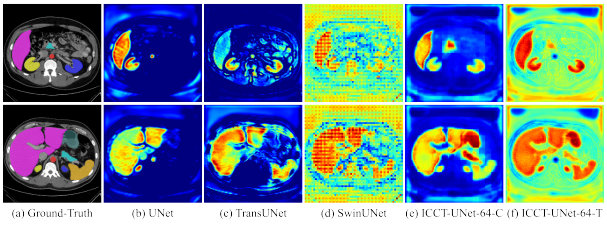

Fig. 5. Visualization of the ground-truth and the feature maps extracted from two *Synapse* [63] images using the decoder of five networks.

method, SwinUNet [46], could capture long-range dependencies, this method usually produced coarse results because it did not extract local structures well. In contrast, the hybrid method, TransUNet [26], achieved the better performance by combining CNNs and Transformers. However, it was inferior to our method due to its serial structure.

In Fig. 5, we further visualize the feature maps extracted using five networks. As can be seen, the ICCT-UNet activated the foreground region better than its counterparts while ignoring the background. These findings should be due to the effective integration of the local features and long-range dependencies.

VI. ABLATION STUDIES

We also conducted extensive experiments to examine the effect of different factors on the performance of our method. For simplicity, only the *Synapse* [63] data set was used.

A. Effect of the Training Setup

To examine the effect of the training setup, we re-trained our network and seven baselines, including UNet [9], XNet [12], TransUNet [26], SwinUNet [46], MissFormer [27], ScaleFormer [52] and UNet-2022 [29], using our training setup for five times. The results are reported in Table X. As can be seen, the baselines normally produced the worse result using our training setup than that obtained using the original setup shown in Table IV. Also, they were outperformed by the proposed method. On the other hand, the DSC values derived using our method and UNet-2022 were 82.32 and 81.43 respectively, when they were trained and tested using the training/inference setup of the UNet-2022. It is likely that our simple training setup accounted for the fact that the proposed method did not always produce the best result. However, we only paid attention to validating the effectiveness of the coupling of CNNs and Transformers for image segmentation rather than deriving the best performance by improving the training setup.

B. Effects of the Dual-Stream Structure and the W-MHCA Unit

To validate the effectiveness of our dual-stream structure and W-MHCA unit, we built three additional feature exchange units by removing the W-MHCA unit or replacing it with the addition or concatenation fusion methods. To overcome the semantic gap between the two streams, the FCU [31] was incorporated into the fusion-based units. Given that the four units were applied to the encoder and decoder of the ICCT-UNet with different combinations, the results obtained using each stream are shown in Table IX.

When the W-MHCA unit was removed from the encoder and decoder, in essence, two independent single-stream networks were created. As can be observed, each of these produced the worst result. It can also be seen that the feature exchange in the encoder always boosted the performance of each stream no matter what feature exchange unit was applied. Given that the dual-stream encoder was utilized, the use of the double-stream decoder improved the performance when the addition fusion or W-MHCA units were used. However, the best performance was produced when our W-MHCA unit was used for both the encoder and decoder. It was indicated that the proposed dual-stream structure and the W-MHCA unit play important roles in image segmentation.

It has been demonstrated that Transformer-based methods can learn the locality when trained with the sufficient data and epochs [49], [71]. As shown in Table IX, the Transformer stream achieved the DSC value of 69.27 after it had been trained for 450 epochs. To investigate whether or not the poor result was due to the insufficient training, we trained the stream for 1000 epochs and the DSC value raised to 74.0. However, this result was still inferior to that of our dual-stream model. By coupling the Transformer stream with a CNN stream, we not only boosted the performance of the Transformer stream but also shortened the training time. Thus, the effectiveness of the dual-stream structure was further indicated.

C. Effect of the Residual Fusion Module

For the purpose of fusing the features extracted at the two streams, we designed a Residual Fusion Module (RFM). To examine the effectiveness and necessity of this module, we replaced it by a convolutional layer. The modified network was re-trained and the DSC values obtained using CNNs, Transformers and the convolution fusion module were 83.95, 82.28 and 83.99 respectively. It was indicated that the application of a fusion approach to the features learned using the two streams has the potential to enhance the segmentation accuracy. Although the convolution fusion module could slightly improve the performance, the DSC value was still lower than 84.60 produced by the RFM. It was suggested that the RFM is effective and necessary for further improving the performance of the proposed ICCT-UNet.

D. Effect of the Model Size

In terms of eight methods, model sizes (i.e., number of parameters), computational complexity (i.e., FLOPs) and the average DSC values derived are compared in Table X. As can be seen, our model outperformed its counterparts with different sizes. Although a larger model may be useful for achieving the better performance, the architecture of the network also plays an important role. To investigate the effect of the model size, we built a variant of the UNet [9] by enlarging the model size to 69.1M. The average DSC value obtained using this model was 80.11 ± 0.65, which was much worse than the value of 84.25 ± 0.75 derived using our method. It was indicated that the superiority of our method should be due to the effective coupling of CNNs and Transformers rather than its model size. Although our method had the highest FLOPs, the inference time that it took was approximately 0.02 seconds per 224×224 image. In this case, our method achieved a proper trade-off between the computational complexity and effectiveness.

E. Effect of the Scale of the Training Set

To examine the effect of the scale of the training set, we re-trained the UNet [9] and our method using 1/4 and 1/2 of the *Synapse* [63] training set. The results are shown in Table XI. Compared with the results shown in Table IV, it can be seen that (1) our ICCT-UNet-F model consistently outperformed the UNet under the two scales, which suggested the generalization and effectiveness of our method; (2) while the UNet achieved comparable results using the smaller training set, the gap between the results of the ICCT-UNet and UNet was large when the larger training set was used. It was indicated that CNNs can be trained well on a relatively small data set but they are hard to gain the greater improvement due to their locality nature, while coupling Transformers to these was useful for boosting their performance; (3) it was difficult to obtain satisfying results using the Transformer stream when our method was trained on a small data set, which hindered the training of the CNN stream. However, the RFM was still helpful for achieving the promising result in this situation.

TABLE IX
COMPARISON OF DIFFERENT FEATURE EXCHANGE UNITS, INCLUDING NONUSE (NONE), THE ADDITION FUSION (ADD), THE CONCATENATION FUSION (CAT) AND W-MHCA, ON THE *Synapse* [63] DATA SET.

Encoder	Decoder	DSC↑		Encoder	Decoder	DSC↑	
		CNN	Trans			CNN	Trans
None	None	81.87	69.27	None	W-MHCA	82.12	81.34
None	Add	82.58	81.50	W-MHCA	None	82.81	79.39
Add	None	82.77	79.59	Add	Add	83.68	81.08
None	Cat	82.01	80.10	Cat	Cat	82.62	80.41
Cat	None	82.99	80.97	W-MHCA	W-MHCA	84.18	83.50

TABLE X
COMPARISON OF THE NUMBER OF PARAMETERS, COMPUTATIONAL COMPLEXITY (I.E., FLOPS) AND AVERAGE DSC VALUES DERIVED USING DIFFERENT METHODS ON THE *Synapse* [63] DATA SET. THE MODELS WERE TRAINED USING OUR SETUP FOR 5 TIMES.

Model	#Params	FLOPs	DSC	Model	#Params	Flops	DSC
UNet	17.2 M	30.7 G	80.51±0.64	MissFormer	35.4 M	7.2 G	77.81±0.56
XNet	34.2 M	52.1 G	81.87±0.96	ScaleFormer	113.7 M	48.5 G	78.89±0.28
TransUNet	105.1 M	28.5 G	77.15±1.18	UNet-2022	72.9 M	10.6 G	79.16±1.17
SwinUNet	41.3 M	8.7 G	72.59±2.13	ICCT-UNet-64	67.4 M	90.9 G	84.25±0.75

TABLE XI
EFFECT OF THE SCALE OF THE TRAINING SET ON THE UNET [9] AND ICCT-UNET WHEN THEY WERE TRAINED USING THE *Synapse* [63] DATA.

Scale of Data Set	UNet	ICCT-UNet-C	ICCT-UNet-T	ICCT-UNet-F
1/4 Training Set	74.94±0.98	68.85±1.56	48.17±8.50	75.56±0.74
1/2 Training Set	77.52±0.48	79.26±2.07	74.63±3.89	81.07±1.19

VII. CONCLUSION

To fulfill the small sample image segmentation task, we proposed a novel network, which comprised an Interactive Coupling of Convolutions and Transformers Based UNet (ICCT-UNet) and a Residual Fusion Module (RFM). Compared with the serial and interleaved networks which simplistically stacked CNNs and Transforms and the disentangled networks which unidirectionally fused CNNs and Transforms, the ICCT-UNet was built on top of a parallel dual-stream architecture, which not only kept each stream relatively independent but also enabled these to interactively exchange features. We also developed a Window-Based Multi-head Cross-Attention (W-MHCA) mechanism. In contrast to the original cross-attention method, the W-MHCA can perform feature exchange at all resolutions with an acceptable computational cost and reduce redundant features. The RFM was used to predict a logit map by fusing the features extracted in the two streams. To our knowledge, both the dual-stream network and the W-MHCA unit have not been applied to image segmentation before.

Experimental results showed that our method performed better than, or at least comparably to, the baselines (including not only the CNN-based and Transformer-based networks but also the serial, interleaved and disentangled hybrid networks) on eight data sets. Either the CNN or the Transformer streams normally outperformed their single-stream counterparts. Also, the result obtained using the RFM was usually better than that produced by a single stream. It was suggested that a complementary action of both the streams has been achieved by our method. We believe that these promising results are due to the inherent capability of our method to effectively integrate the local structure extracted using the CNN stream and the global context captured using the Transformer stream, achieved through the proposed W-MHCA mechanism.

It is noteworthy that we mainly paid attention to investigating the effectiveness of the iterative coupling of CNNs and Transformers for image segmentation. This explains why we only used a simply training setup which was much simpler than that used by the state-of-the-art methods, e.g., UNet-2022. However, the key point is that we have shown that coupling CNNs and Transformers is useful for image segmentation.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2015.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.

[4] L. Salgado, N. Garcia, J. Menendez, and E. Rendon, "Efficient image segmentation for region-based motion estimation and compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 7, pp. 1029–1039, Oct. 2000.

[5] M. S. Allili, D. Ziou, N. Bouguila, and S. Boutemedjet, "Image and video segmentation by combining unsupervised generalized gaussian mixture modeling and feature selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 10, pp. 1373–1377, Oct. 2010.

[6] X. Sun, C. Chen, X. Wang, J. Dong, H. Zhou, and S. Chen, "Gaussian dynamic convolution for efficient single-image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2937–2948, May 2022.

[7] J. Ji, R. Shi, S. Li, P. Chen, and Q. Miao, "Encoder-decoder with cascaded crfs for semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1926–1938, May 2021.

[8] X. Dong, C. J. Taylor, and T. F. Cootes, "Automatic aerospace weld inspection using unsupervised local deep feature learning," *Knowl Based Syst.*, Jun. 2021.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 2015, pp. 234–241.

[10] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," *2018 Int. Conf. Inf. Technol. Med. Edu.*, pp. 327–331, 2018.

[11] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Trans. Med. Imag.*, vol. 37, pp. 2663–2674, Dec. 2018.

[12] 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.

[13] H. Li, D.-H. Zhai, and Y. Xia, "Erdunet: An efficient residual double-coding unet for medical image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2023.

[14] G. Bhattacharya, B. Mandal, and N. B. Puhan, "Multi-deformation aware attention learning for concrete structural defect classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3707–3713, Sept. 2021.

[15] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "Deepcrack: Learning hierarchical convolutional features for crack detection," *IEEE Trans. Image Process.*, vol. 28, pp. 1498–1512, Mar. 2019.

[16] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[21] L. Attard, C. J. Debono, G. Valentino, and M. di Castro, "Tunnel inspection using photogrammetric techniques and image processing: A review," *ISPRS J. Photogramm. Remote Sens.*, 2018.

[22] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. Á. G. Ballester, G. Sanroma, S. Napel, S. E. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. F. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Igum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[23] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2019.

[24] 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," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 9992–10 002, 2021.

[25] H. Wu, B. Xiao, N. C. F. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 22–31, 2021.

[26] 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*, vol. abs/2102.04306, 2021.

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

[28] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[29] J. Guo, H.-Y. Zhou, L. Wang, and Y. Yu, "Unet-2022: Exploring dynamics in non-isomorphic architecture," *arXiv preprint arXiv:2210.15566*, 2022.

[30] N. Park and S. Kim, "How do vision transformers work?" *arXiv preprint arXiv:2202.06709*, 2022.

[31] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 357–366, 2021.

[32] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-former: Bridging mobilenet and transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 5270–5279.

[33] G. Asaeikheybari, J. Green, X. Qian, H. Jiang, and M.-C. Huang, "Medical image learning from a few/few training samples: Melanoma segmentation study," *Smart Health*, vol. 14, p. 100088, 2019.

[34] N. Ibtehaz and M. S. Rahman, "Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks : the official journal of the International Neural Network Society*, vol. 121, pp. 74–87, 2020.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1–9, 2015.

[36] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, vol. 11045, pp. 3–11, 2018.

[37] D. Jha, M. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," *2020 IEEE 33rd Int. Symp. Comput.-Based Med. System. (CBMS)*, pp. 558–564, 2020.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

[39] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "Dunet: A deformable network for retinal vessel segmentation," *Knowl. Based Syst.*, vol. 178, pp. 149–162, Aug. 2019.

[40] X. Dong, C. J. Taylor, and T. F. Cootes, "Small defect detection using convolutional neural network features and random forests," in *Proc. Eur. Conf. Comput. Vis. Workshop.*, 2018.

[41] Y. Fei, K. C. P. Wang, A. Zhang, C. Chen, J. Q. Li, Y. Liu, G. Yang, and B. Li, "Pixel-level cracking detection on 3d asphalt pavement images through deep-learning- based cracknet-v," *IEEE Trans. Intell. Transp. Syst.*, Jan. 2020.

[42] X. Dong, C. J. Taylor, and T. F. Cootes, "Defect detection and classification by training a generic convolutional neural network encoder," *IEEE Trans. Signal Process.*, vol. 68, pp. 6055–6069, 2020.

[43] X. Dong, C. J. Taylor, and T. Cootes, "Defect classification and detection using a multitask deep one-class cnn," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, pp. 1719–1730, July 2022.

[44] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12 077–12 090, 2021.

[45] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6877–6886, 2020.

[46] 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 *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 205–218.

[47] H. Liu, X. Miao, C. Mertz, C. Xu, and H. Kong, "Crackformer: Transformer network for fine-grained crack detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 3783–3792.

[48] J. Chen, N. Zhao, R. Zhang, L. Chen, K. Huang, and Z. Qiu, "Refined crack detection via lecsformer for autonomous road inspection vehicles," *IEEE trans. intell. veh.*, vol. 8, no. 3, pp. 2049–2061, Mar. 2023.

[49] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks," *Proc. Adv. Neural Inf. Process. Syst.*, 2021.

[50] Y. Gao, M. Zhou, and D. N. Metaxas, "Utnet: a hybrid transformer architecture for medical image segmentation," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*. Springer, 2021, pp. 61–71.

[51] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo, "Multi-compound transformer for accurate biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*. Springer, 2021, pp. 326–336.

[52] H. Huang, S. Xie, L. Lin, Y. Iwamoto, X. Han, Y.-W. Chen, and R. Tong, "Scaleformer: Revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation," *arXiv preprint arXiv:2207.14552*, 2022.

[53] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *2016 Int. Conf. 3D Vis.*, pp. 565–571, 2016.

[54] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*. Springer, 2016, pp. 424–432.

[55] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[56] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[57] D. Karimi, S. D. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using transformers," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 2021.

[58] A. Hatamizadeh, D. Yang, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 1748–1758, 2022.

[59] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Int. MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[60] C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, and S. Yan, "Inception transformer," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 23 495–23 509, 2022.

[61] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 648–656, 2015.
[62] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021.
[63] 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, 2015, p. 12.
[64] N. C. F. Codella, D. A. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. C. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi)," *Int. Symp. Biomed. Imaging*, 2018.
[65] W. Al-Dhabyani, M. M. M. Gomaa, H. Khaled, and A. A. Fahmy, "Dataset of breast ultrasound images." *Data in Brief*, Feb. 2020.
[66] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. H. Menze, O. Ronneberger, R. M. Summers, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, J. Golia-Pernicka, S. Heckers, W. R. Jarnagin, M. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, and M. J. Cardoso, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.
[67] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Trans. Intell. Transp. Syst.*, Dec. 2016.
[68] Y. Huang, C. Qiu, Y. Guo, X. Wang, and K. Yuan, "Surface defect saliency of magnetic tile," *Conf. Autom. Sci. Eng.*, 2018.
[69] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-Based Deep-Learning Approach for Surface-Defect Detection," *J. Intell. Manuf.*, May 2019.
[70] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[71] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," *arXiv preprint arXiv:1911.03584*, 2019.

Hao Qi received the bachelor's degree in Management from the Ocean University of China (OUC), Qingdao, Shandong Province, China, in 2020. He is currently a post-graduate student at Ocean University of China working toward his master's degree in Computer Science. His research interests include computer vision, deep learning, image segmentation, and image enhancement.

Huiyu Zhou received the B.Eng. degree in radio technology from the Huazhong University of Science and Technology, Wuhan, China, in 1990, the M.Sc. degree in biomedical engineering from the University of Dundee, Dundee, U.K., in 2002, and the Ph.D. degree in computer vision from Heriot-Watt University, Edinburgh, U.K., in 2006. He is currently a Full Professor with the School of Computing and Mathematical Sciences, University of Leicester, Leicester, U.K. His research interests include medical image processing, computer vision, intelligent systems, and data mining.

Junyu Dong received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, U.K., in 2003. He joined Ocean University of China in 2004. He is currently a Professor and the Dean of the Faculty of Information Science and Engineering, Ocean University of China. His research interests include computer vision, underwater image processing, and machine learning, with more than ten research projects supported by the NSFC, MOST, and other funding agencies.

Xinghui Dong received the PhD degree from Heriot-Watt University, U.K., in 2014. He worked with the Centre for Imaging Sciences, the University of Manchester, U.K., between 2015 and 2021. Then he jointed Ocean University of China in 2021. He is currently a professor at the Ocean University of China. His research interests include computer vision, defect detection, texture analysis, and visual perception.