TransAttUnet: Multi-Level Attention-Guided U-Net With Transformer for Medical Image Segmentation

Bingzhi Chen, Yishu Liu, Zheng Zhang, and Guangming Lu, Senior Member, IEEE, and Adams Wai Kin Kong, Member, IEEE

Abstract—Accurate segmentation of organs or lesions from medical images is crucial for reliable diagnosis of diseases and organ morphometry. In recent years, convolutional encoder-decoder solutions have achieved substantial progress in the field of automatic medical image segmentation. Due to the inherent bias in the convolution operations, prior models mainly focus on local visual cues formed by the neighboring pixels, but fail to fully model the long-range contextual dependencies. In this article, we propose a novel Transformer-based Attention Guided Network called TransAttUnet, in which the multi-level guided attention and multi-scale skip connection are designed to jointly enhance the performance of the semantical segmentation architecture. Inspired by Transformer, the self-aware attention (SAA) module with Transformer Self Attention (TSA) and Global Spatial Attention (GSA) is incorporated into TransAttUnet to effectively learn the non-local interactions among encoder features. Moreover, we also use additional multi-scale skip connections between decoder blocks to aggregate the upsampled features with different semantic scales. In this way, the representation ability of multi-scale context information is strengthened to generate discriminative features. Benefiting from these complementary components, the proposed TransAttUnet can effectively alleviate the loss of fine details caused by the stacking of convolution layers and the consecutive sampling operations, finally improving the segmentation quality of medical images. Extensive experiments were conducted on multiple medical image segmentation datasets from various imaging modalities, which demonstrate that the proposed method consistently outperforms the existing state-of-the-art methods.

Index Terms—Medical image segmentation, transformer, multi-level guided attention, multi-scale skip connection.

I. INTRODUCTION

In clinical diagnosis, the core of medical image segmentation is to delineate the objects of interest from the complex background on various biomedical images [1], [2], [3], such as X-ray, Computerized Tomography (CT), Magnetic Resonance Imaging (MRI), and Ultrasound. It is useful for quantitative diagnosis and morphological analysis of specific lesions in human organs and tissues. As shown in Fig. 1, it requires enormous effort and patience to handle the complex contours and textures. However, traditional manual annotation heavily relies on clinical experiences. Measurements based on manual annotations by clinicians might be highly accurate, but they can also be labor-intensive under typical clinical settings. Therefore, it is in great demand to develop accurate medical image segmentation methods.

In the field of smart medicine, deep encoder-decoder architectures have gradually become the facto benchmark for medical image segmentation owing to their demonstrated effectiveness in recovering object details. The main purpose of convolutional operations used in deep encoder-decoder architectures is to extract local features of images by gathering local information from the neighboring pixels. Typically, the stacks of convolution layers and the consecutive sampling operations constantly extend the receptive field and aggregate the global filter responses to determine the coarse object boundaries. Benefiting from deconvolution operations [4], Fully Convolutional Networks (FCN) [5] can naturally operate on images with various sizes and generate appropriate dimension outputs according to the corresponding inputs. Inspired by FCN, U-Net [6] directly combines low-level features from the analysis path and deep features in the expansion path through encoder-decoder skip connections, which can achieve the trade-off capability between local information and contextual information.

Despite the interesting design and encouraging performance, the above design limits the information flow due to the convolutional operations and becomes a bottleneck for performance enhancement. Specifically, these architectures have several drawbacks: 1) the inherent inductive bias of convolution computing paradigm typically limit the architectures’ ability to model long-range feature dependencies, since their convolutional operations are restricted to local neighborhoods of the input; 2) the low-level features might be prevented from being transmitted to the subsequent convolutional layers in the process of pooling and convolution [8], which compromise the quality
of local information and degrade the semantic segmentation performance; 3) the existing skip connection mechanism is only performed on the same scale feature maps without exploring the relationship between feature maps from different decoding stages, which cannot guarantee the consistency of feature representations and semantic embeddings [9].

To solve these issues, researchers have put considerable effort in developing various variants of U-Net. The existing works follow three research lines: 1) attention-guided approaches; 2) context-based approaches; 3) Transformer-based approaches. For example, prior attention-guided works, e.g. Attention U-Net [10] and Channel-U-Net [8], attempted to leverage various attention mechanisms to optimize the spatial information of feature maps extracted from encoding, decoding, and output stages. The main idea of these attention mechanisms is to generate a confidence mask to recalibrate the response of the original feature maps. Different from the above approaches, the second one is to explore the contextual information by using multi-scale connections. It is indisputable that both high-level abstract information and low-level pixel information are of great significance in developing accurate segmentation. Therefore, varying degrees of context-based shortcut connection components have been embedded into the U-shaped architectures, such as Unet++ [9], and MA-U-Net [11], in order to capture broader and deeper contextual representations. Recently, Transformer [12], [13] is getting great attention from computer vision (CV) researchers. Especially, some recent works, such as TransUNet [14] and MedT [15], have tried to incorporate Transformer with CNN-based model to boost the performance of medical image segmentation. In general, Transformer-based approaches have very high computation complexities to process feature maps, and often required models pretrained on a large external dataset. As the core of Transformer, the multi-head mechanism [16] can capture the long-range contextual information by running through the scaled dot-product attention multiple times in parallel. Therefore, it is undoubtedly an ideal supplemental part that can efficiently compensate for the design flaws of U-Net.

As mentioned, how to overcome the inherent bias in the convolution operations and effectively model the long-range contextual dependencies for the standard U-Net are of great importance for medical image segmentation. Inspired by these considerations, in this article, we propose a novel multi-level attention-guided U-Net with Transformer, dubbed TransAttUnet, that can effectively enhance the segmentation accuracy of traditional U-shaped architecture by jointly utilizing multi-level guided attention and multi-scale skip connection. The architecture of the proposed TransAttUnet is shown in Fig. 2. Specifically, a robust self-aware attention (SAA) module is first embedded into the proposed TransAttUnet as the bridge between the encoder and decoder subnetworks. As the key component of TransAttUnet, the purpose of the SAA module is to concurrently leverage the powerful abilities of transformer self attention (TSA) and global spatial attention (GSA) to establish effective long-range interactions and global spatial relationships between encoder semantic features. Motivated by the idea of residual and dense shortcut connections, a multi-scale skip connection scheme is added into decoder sub-networks with a series of transition operations, including upsampling, concatenation, and convolution. Thus, it can flexibly aggregate the residual or dense contextual feature maps from decoder blocks of varying semantic scales step by step, in order to generate more discriminative feature representations. With the contributions of these complementary components, the proposed method can achieve accurate semantic segmentation masks of medical images. We evaluate the effectiveness of our proposed TransAttUnet on multiple medical image datasets acquired from various imaging modalities. Our experiments focus on addressing five typical challenges in clinical diagnosis, which include: 1) skin lesion segmentation on dermatoscopic images [17]; 2) lung segmentation on chest X-ray images; 3) COVID-19 pneumonia lesion segmentation on chest CT images; 4) nuclei segmentation on divergent images [18]; and 5) gland segmentation on histology images [19]. Our main contributions are summarized as follows:

- This article proposes a Transformer-Attention based U-shaped framework (TransAttUnet), that integrates the advantages of multi-level guided attention and multi-scale skip connections into the standard U-Net to improve segmentation performance for various medical images.
- By combining transformer self-attention with global spatial attention, our TransAttUnet can effectively model both contextual semantic information and global spatial relationships, ensuring the consistency of feature representations and semantic embeddings.
- Compared to the one-step cascade connection, our proposed residual or dense step-growth connections can not only reduce the noise disturbance but also mitigate the loss of fine details that may occur when directly upsampling with large scales.
- Extensive experimental results on the five medical image datasets demonstrate the superiorities and generalizability of the proposed TransAttUnet for automatic medical image segmentation in comparison with state-of-the-art baselines.

II. RELATED WORK

In this section, we will provide a brief overview of two related topics. The first topic pertains to the state-of-the-art variants of U-Net used for automatic medical image segmentation, while
the second topic focuses on the applications of the Transformer in computer vision.

A. Variants of U-Net

Many efforts have been devoted to optimizing the structure of the U-Net in the field of automatic medical image segmentation. Typically, these variants can be roughly divided into two categories, i.e., attention-guided approaches and multi-scale context approaches.

1) Attention Guided Approaches: Various attention-guided methods have been proposed to achieve accurate segmentation of objects in medical images captured using varying imaging modalities. One such method is Attention U-Net [10], which uses an attention gate (AG) to suppress irrelevant feature responses and emphasize salient features, thereby enhancing the feature learning ability of U-Net. This improves the model’s sensitivity and prediction accuracy. Channel-UNet [8] introduces the spatial channel convolution to determine the optimal mapping relationship among spatial information from different patches. By simultaneously recalibrating the different types of features at the spatial and channel levels, SCAU-Net [20] can guide the model to neglect irrelevant information and focus on more discriminant regions of the image. Similarly, 3D attention U-Net [21] applied the channel and spatial attention into the decoder subnetwork of 3D U-Net [22] to segment brain tumors in MRI images. XLSor [23] makes use of the criss-cross attention module to aggregate long-range pixel-wise contextual information in both horizontal and vertical directions for lung segmentation in chest X-ray images. Additionally, Residual Attention U-Net [24] leverages the soft attention mechanism to enhance the model’s ability to differentiate between a range of COVID-19 symptoms in chest CT images. However, the direct application of the attention mechanism might result in the loss of available feature representations, especially when the judgment for the region of interest goes wrong, which is inadequate to meet the needs of an ideal model.

2) Multi-Scale Context Approaches: To make full use of the multi-scale context information, prior works have attempted to combine the low-level features from the shallow layers with the high-level feature of the deep layers for retaining the detailed image information. For example, UNet++ [9] proposes a highly flexible multi-scale feature fusion scheme by aggregate features of varying semantic scales with redesigned skip connections. U2-Net [25] exploits Residual U-blocks to capture intra-stage contextual information and directly cascades the inter-block feature mapping of the decoder subnetwork to mitigate the loss of fine details. MA-Unet [11] establishes a multi-scale mechanism to directly aggregating global contextual information of different scales from intermediate layers as final feature representations, and it also utilizes additional attention mechanisms to improve the prediction accuracy for medical image segmentation. However, the one-step cascade connection method may overlook crucial details during the large-scale upsampling process, resulting in the loss of valuable information.

B. Transformer in CV

1) Transformer for Various Vision Tasks: The increasing use of the Transformer in various NLP tasks has led to the development of several Transformer-based methods for computer vision tasks. In particular, ViT [26] presents a pure self-attention Vision Transformer for image recognition, which is the first attempt of the transformer-based method to surpass the traditional CNN-based works. By introducing Transformer into CNNs, DETR [27] proposes a fully end-to-end object detector to eliminate the hand-designed components in CNN-based object detectors. Subsequently, SETR [28] replaces the encoder in the standard FCN architecture with Transformer and achieves encouraging performance on the natural image segmentation task.

2) Transformer for Medical Image Segmentation: TransUNet [14] draws inspiration from ViT and employs the Transformer as an encoder, which is applied with U-Net to enhance the performance of medical image segmentation tasks. On the
TABLE I
NECESSARY NOTATIONS AND THE CORRESPONDING DEFINITIONS

| Notation | Description |
|---------|-------------|
| $X$     | The input medical image |
| $F$     | The encoder features |
| $F_{\text{TSA}}$ | The output feature of the TSA component |
| $F_{\text{GSA}}$ | The output feature of the GSA component |
| $F_{\text{SAA}}$ | The output feature of the SAA module |
| $F_t$  | The output feature of multi-scale skip connection |
| $v_{\text{up}}(\cdot)$ | The upsampling operation in nth stage |
| $f_{\text{conv}}(\cdot)$ | The mixed convolution operation in nth stage |
| $\mathcal{L}_{\text{BCE}}$ | Binary Cross-Entropy loss |
| $\mathcal{L}_{\text{Dice}}$ | Sorensen-Dice loss |
| $\mathcal{L}$ | The total objective function |

other hand, Moreover, TransFuse [29] uses a parallel approach to combine Transformer and CNN for modeling global context information, which improves the efficiency of the process. MC-Trans [30] leverages the Transformer to incorporate rich context modeling and semantic relationship mining for precise medical image segmentation. Additionally, MedT [15] introduces a Gated Axial-Attention model that employs a Transformer-based gated position-sensitive axial attention mechanism for medical image segmentation. In contrast, the proposed TransAttUnet aims to explore the potential of using the Transformer to address U-Net’s inability to model long-range contextual interactions.

III. METHODOLOGY

This section introduces the proposed TransAttUnet, which comprises several components. Firstly, we provide an overview of TransAttUnet. Next, we describe the principles and structure of TransAttUnet in detail, along with an explanation of each component. We also present the unified loss function used in our TransAttUnet. Additionally, Table I provides the necessary notations and the corresponding definitions for reference.

A. Overview of TransAttUnet

Denote the input medical image as $X \in \mathbb{R}^{C \times H \times W}$, where $C$ is the number of channels, and $H \times W$ represents the spatial resolution of the image instance. The goal of this work is to automatically segment medical images and generate pixel-wise semantic label maps of size $H \times W$. The general learning framework is illustrated in Fig. 2. As with the previous works, the proposed TransAttUnet is also built on the standard encoder-decoder U-shaped architecture, as shown in Fig. 2. To overcome the limitations mentioned in Section I, TransAttUnet aims to leverage multi-level complementary self-aware attention components, as well as multi-scale skip connections [9], [11] to further improve the semantic segmentation quality of medical images. Compared to the standard U-Net, the SAA module in TransAttUnet leverages the benefits of both TSA and GSA mechanisms to perceive long-range contextual information, thereby enhancing the representation ability of the encoder’s semantic features. Furthermore, the multi-scale skip connections used in TransAttUnet are designed to achieve the residual and dense shortcut connections between the intermediate layers of different semantic scales, which can aggregate the contextual information for multi-scale prediction fusion.

B. Self-Aware Attention Module

The proposed TransAttUnet differs from previous methods by incorporating a self-aware attention module between the encoder and decoder subnetworks. This module contains two independent self-attention mechanisms: transformer self-attention (TSA) and global spatial attention (GSA). By utilizing these mechanisms, TransAttUnet can capture wider and richer contextual representations compared to the standard U-Net. The module is positioned at the bottom of the U-shaped architecture, which enables the module to act as a bridge between the encoder and decoder, enhancing the model’s ability to capture long-range contextual information.

1) Transformer Self Attention: Our TransAttUnet first introduces a TSA component that leverages the multi-head self-attention function from Transformer to capture semantic information from global representation subspaces. To incorporate information on absolute and relative position, the TSA component employs learned positional encoding that can be shared across all attention layers for a given query/key-value sequence. The multi-head attention mechanism processes each attention head separately and combines them through another self-attention function from Transformer to capture semantic contextual information.

2) Global Spatial Attention: The SAA module incorporates the GSA component to enhance the learned features with global context and encode broader contextual positional information into local features, thereby improving the intra-class compactness and optimizing the feature representations. As shown in...
Fig. 2, the architecture of the GSA component is illustrated by the blue region.

To begin with, two types of convolutional operations are applied to the encoder features $F_{en}$ to generate two feature maps: $F_p^c \in \mathbb{R}^{c \times h \times w}$ and $F_p^s \in \mathbb{R}^{c \times h \times w}$, where $c' = c/8$. Subsequently, $F_p^c$ is reshaped and transposed into feature maps $M \in \mathbb{R}^{(h \times w) \times c'}$ and $N \in \mathbb{R}^{c' \times (h \times w)}$, while $F_p^s$ is transposed in to $W \in \mathbb{R}^{c' \times (h \times w)}$, respectively. Next, a matrix multiplication operation is performed between $M$ and $N$, followed by a softmax normalization to obtain the position attention maps $B \in \mathbb{R}^{(h \times w) \times (h \times w)}$. The equation for computing $B$ is as follows:

$$B_{i,j} = \frac{\exp(M_i \cdot N_j)}{\sum_{j=1}^{M} \exp(M_i \cdot N_j)},$$

where $B_{i,j}$ represents the influence of the $i$-th position on the $j$-th position, $n = h \times w$ denotes the total number of pixels in the feature map. After computing the position attention maps $B$, $W$ is multiplied with $B$, and the resulting feature at each position can be formulated as:

$$GSA(M, N, W)_p = \sum_{q=1}^{h \times w} (W_q B_{p,q}).$$

Moreover, we reshape the resulting features to obtain the final output of GSA, which is denoted as $F_{gsa} \in \mathbb{R}^{c \times h \times w}.

3) Attention Embedding Fusion: To make full use of the obtained contextual information and spatial relationships, a weighted combination scheme for the original and attention feature embeddings is used at the end of the SSA module, which is defined as:

$$F_{SA} = \lambda_1 F_{tsa} + \lambda_2 F_{gsa} + F_{en},$$

where $\lambda_1$ and $\lambda_2$ are the scale parameters that control the importance of the self attention maps and spatial attention maps, respectively. The weights of both components are initialized to 0 and are progressively increased to assign greater importance to the relevant features, thereby enhancing the feature representations with semantic consistency.

C. Multi-Scale Skip Connection

Notably, many advanced works [11], [14], [25] have demonstrated the effectiveness of multi-scale feature fusion in encoding global and local contexts. Specifically, the multi-scale skip connection scheme aims to aggregate the features of varying semantic scales with a series of transition operations, including upsampling, concatenation, and convolution. Inspired by previous works, three different types of connections, Cascade Connection, Residual Connection, and Dense Connection shown in Fig. 3 are investigated in this study.

1) Cascade Connection: The feature maps from all blocks with varying semantic scales are upsampled using bilinear interpolation to a common resolution and concatenated to form a unified feature representation. The formulation for this is:

$$F = f_n(v_1 F_1) \oplus v_2 F_2 \oplus \cdots \oplus F_n, \quad (6)$$

where $\oplus$ denotes concatenation operations, $v_n(\cdot)$ and $f_n(\cdot)$ are the upsampling and mixed convolution operations in $n$th stage, respectively.

2) Residual Connection: To incorporate residual connections into each decoder block, the input feature maps are first up-sampled to the same resolution as the output feature maps through bilinear interpolation. Then, the up-sampled input feature maps are concatenated with the output feature maps, and the resulting feature maps are used as inputs for the subsequent blocks. This process is formulated as follows:

$$F_n = f_n((F_n) \oplus v_{n-1} F_{n-1}). \quad (7)$$

3) Dense Connection: The upsampling features of previous decoder blocks are integrated as the inputs of the current block, and the output feature maps are used as inputs into all subsequent blocks, which is formulated as:

$$F_n = f_n(v_1 F_1) \oplus v_2 F_2 \oplus \cdots \oplus v_{n-1} F_{n-1}). \quad (8)$$

In particular, the proposed TransAttUnet focuses on two different multi-scale skip connection schemes, i.e., the residual connection and dense connection, to guide the upsampling process in the decoder subnetwork. Compared to the existing works that only using the one-off cascade connection, the residual or dense step-growth connections can gradually aggregate multiple decoder features of varying semantic scales to generate the most discriminative feature representations. In this way, the proposed TransAttUnet not only can mitigate the loss of fine details caused by over-uptampling, but also alleviate the problems of vanishing-gradient and overfitting.
D. Training and Optimization

During the training process, the TransAttUnet model is trained end-to-end using an objective function. The objective function is computed using the Sorensen-Dice loss and Binary Cross-Entropy function. The pixel-wise soft-max over the final feature maps is used in the calculation of the objective function. The objective function can be formulated as follows:

\[
\mathcal{L}_{BCE} = \frac{1}{t} \sum_{i=1}^{t} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)),
\]

\[
\mathcal{L}_{Dice} = 1 - \frac{\sum_{i=1}^{t} y_i p_i + \varepsilon}{\sum_{i=1}^{t} y_i + p_i + \varepsilon},
\]

\[
\mathcal{L} = \alpha \cdot \mathcal{L}_{BCE} + \beta \cdot \mathcal{L}_{Dice},
\]

where \( t \) is the total number of pixels in each image, \( y_i \) represents the ground-truth value of the \( i \)th pixel, and \( p_i \) is the confidence score of the \( i \)th pixel in prediction results. Moreover, \( \alpha \) and \( \beta \) represent the weights of the Binary Cross-Entropy loss and Sorensen-Dice loss, respectively. In our experiment, \( \alpha = \beta = 0.5 \), and \( \varepsilon = 10^{-6} \).

IV. EXPERIMENTS

In this section, we evaluate the performances of the proposed TransAttUnet framework on multiple benchmark datasets by comparing it with the state-of-the-art baselines. Next, we make a detailed discussion for the ablation studies. Finally, visualization analysis of decoder stages is presented.

A. Datasets

To verify the effectiveness and efficiency of our TransAttUnet, we first conducted comparative experiments for the task of skin lesion segmentation on ISIC-2018 [31] dataset, and lung field segmentation on the combination of the JSRT [32], Montgomery [33], and NIH [23] datasets. Moreover, the proposed TransAttUnet is also evaluated on the Clean-CC-CCII dataset [34], 2018 Data Science Bowl (Bowl) dataset [35], and the Gland Segmentation (GlaS) dataset [19]. Typically, the images from the CC-CCII, Bowl, and GLAS datasets might contain multiple segmenting objects with varying sizes and textures, which greatly increases the segmentation difficulty and complexity.

1) ISIC-2018: The ISIC-2018 is a large-scale dataset of dermoscopy images provided for ISIC-2018 challenge [31] and contains three tasks, including lesion segmentation, lesion attribute detection, and disease classification. Notably, this article focuses on the task of lesion segmentation from dermoscopic images by various types of dermoscopy. In the experiments, 2596 images with the corresponding annotations are selected for experiments and they are randomly divided into 2076 images to form the training set and 520 images as the testing set.

2) JSRT, Montgomery & NIH: Three datasets of frontal chest X-ray images, JSRT, Montgomery, and NIH are involved in our experiments for automatic lung field segmentation. The JSRT dataset comprises 247 chest X-ray images, among which 154 images are abnormal with pulmonary nodule and 93 images are normal. By contrast, the Montgomery dataset contains 138 chest X-ray images, including 80 normal patients and 58 patients with manifested tuberculosis. Different from the JSRT and Montgomery, the NIH dataset contains 178 chest X-ray images with various severity of lung diseases, which can hugely complicate the task of lung field segmentation. Following the settings of previous works [23], we combine these datasets and randomly split them into 407 images for training and 178 images for testing.

3) Clean-CC-CCII: As a large public chest CT dataset for automated COVID-19 diagnosis, the Clean-CC-CCII dataset contains thousands of annotated CT scans from 2,698 patients. In particular, it also provides researchers with high-quality annotation of infection marks to develop a robust model of COVID-19 pneumonia lesion segmentation. In our experiment, 260 CT slices with pixel-wise annotations of COVID-19 pneumonia lesion are selected and randomly split into two subsets, i.e., a training set of 200 images and a test set of 60 images.

4) Bowl: The Bowl dataset is established for the development of robust automatic nucleus segmentation algorithms. It provides participants with a training set of 671 nuclei images along with pixel-wise masks for the nuclei and a test set of 3020 images, which are extracted from 15 diverse image sets of biological experiments. Note that each image contains dozens of nuclei with different sizes. Due to the lack of the annotation masks of the test set, we only evaluate the performance of the proposed method based on the training set. Following the settings of the existing work [36], the training set is split into three subsets: 80% for training, 10% for validation, and 10% for testing.

5) GlaS: The GlaS dataset is a collection of colon histology images from the Colon Histology Images Challenge Contest of MICCAI 2015, which focuses on developing methods for quantifying the morphology of glands. It consists of 165 histology images of stage T3 or T4 colorectal adenocarcinoma, derived from 16 H&E stained histological sections of different patients. One of the challenges of this dataset is the high inter-subject variability in both stain distribution and tissue structure due to processing in the laboratory on different occasions. To evaluate the proposed method, we randomly split the GlaS dataset into training and testing sets of 85 and 80 images, respectively, following the previous works [15], [37].

B. Experimental Settings

1) Baselines: In our experiments, we evaluate three different versions of TransAttUnet, namely TransAttUnet_C, TransAttUnet_D, and TransAttUnet_R. TransAttUnet_C refers to the version where the decoder blocks are connected with one-off cascade connections, as proposed in previous works [11], [14], [25]. TransAttUnet_D uses dense operations to connect the decoder blocks, while TransAttUnet_R uses residual operations. To compare the performance of TransAttUnet, we use the vanilla U-Net [6] as the baseline, as well as three other broad approaches: attention-guided approaches, multi-scale context approaches, and Transformer-based approaches.
• **Attention-guided approaches**:
  Our comparative experiments include three broad approaches as baselines, including attention-guided models such as Attention U-Net [10], Attention R2U-Net [38], Channel-UNet [8], XLSoR [23], FANet [39], and PraNet [40], which have been proposed to improve segmentation performance by incorporating attention mechanisms.

• **Multi-scale context approaches**:
  Meanwhile, several multi-scale context models are used as the major contenders, including Unet++ [9], R2U-Net [38], ResUNet [41], ResUNet++ [42], BCDU-Net [43], KiU-Net [37], and DoubleU-Net [36].

• **Transformer-based approaches**:
  Moreover, various Transformer-based methods, i.e., MedT [15], MCTrans [30], Swin-UNet [44], and SegFormer [45] are regarded as important methods for comparison.

2) **Implementation Details**: To ensure a fair comparison with previous studies, we resize the input images from Clean-CC-CII, JSRT, Montgomery, and NIH datasets to 512 × 512 during both training and testing phases. For the ISIC-2018 and Bowl datasets, we resize the images to 256 × 256. All input images from the GlaS dataset are resized to 128 × 128 uniformly. The TSA module in the proposed TransAttUnet framework is designed to use 8 parallel attention heads. We use stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0001 to optimize the training process. The implementation of the proposed TransAttUnet model is done using PyTorch deep learning toolbox, and all experiments are run on a single Nvidia Titan XP GPU. We train the TransAttUnet framework for 100 epochs with a batch size of 4, and the initial learning rate is set to 0.0001, which decays by a factor of 10 every 40 epochs. During our experiments, we report the probability maps generated directly by the models as the results. The probability maps are binarized using a threshold of 0.5 to obtain binary masks for performance evaluation.

3) **Evaluation Metrics**: In our experimental setup, we employ the mean Dice coefficient (DICE) as the primary evaluation metric for assessing the similarity between the predicted mask and the ground truth. Additionally, we compute four other evaluation metrics, namely mean Intersection over Union (IoU), accuracy (ACC), recall (REC), and precision (PRE) scores, on a per-pixel basis to quantify the segmentation performance. These metrics are computed based on the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values,

\[
\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN},
\]

\[
\text{IoU} = \frac{TP}{TP + FP + FN},
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},
\]

\[
\text{Recall} = \frac{TP}{TP + FN},
\]

\[
\text{Precision} = \frac{TP}{TP + FP},
\]

(C10)

3) **Multi-scale context approaches**: Meanwhile, several multi-scale context models are used as the major contenders, including Unet++ [9], R2U-Net [38] ResUNet [41], ResUNet++ [42], BCDU-Net [43], KiU-Net [37], and DoubleU-Net [36].

4) **Transformer-based approaches**: Moreover, various Transformer-based methods, i.e., MedT [15], MCTrans [30], Swin-UNet [44], and SegFormer [45] are regarded as important methods for comparison.

2) **Implementation Details**: To ensure a fair comparison with previous studies, we resize the input images from Clean-CC-CII, JSRT, Montgomery, and NIH datasets to 512 × 512 during both training and testing phases. For the ISIC-2018 and Bowl datasets, we resize the images to 256 × 256. All input images from the GlaS dataset are resized to 128 × 128 uniformly. The TSA module in the proposed TransAttUnet framework is designed to use 8 parallel attention heads. We use stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0001 to optimize the training process. The implementation of the proposed TransAttUnet model is done using PyTorch deep learning toolbox, and all experiments are run on a single Nvidia Titan XP GPU. We train the TransAttUnet framework for 100 epochs with a batch size of 4, and the initial learning rate is set to 0.0001, which decays by a factor of 10 every 40 epochs. During our experiments, we report the probability maps generated directly by the models as the results. The probability maps are binarized using a threshold of 0.5 to obtain binary masks for performance evaluation.

3) **Evaluation Metrics**: In our experimental setup, we employ the mean Dice coefficient (DICE) as the primary evaluation metric for assessing the similarity between the predicted mask and the ground truth. Additionally, we compute four other evaluation metrics, namely mean Intersection over Union (IoU), accuracy (ACC), recall (REC), and precision (PRE) scores, on a per-pixel basis to quantify the segmentation performance. These metrics are computed based on the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values,

\[
\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN},
\]

\[
\text{IoU} = \frac{TP}{TP + FP + FN},
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},
\]

\[
\text{Recall} = \frac{TP}{TP + FN},
\]

\[
\text{Precision} = \frac{TP}{TP + FP},
\]

(C10)

3) **Multi-scale context approaches**: Meanwhile, several multi-scale context models are used as the major contenders, including Unet++ [9], R2U-Net [38] ResUNet [41], ResUNet++ [42], BCDU-Net [43], KiU-Net [37], and DoubleU-Net [36].

4) **Transformer-based approaches**: Moreover, various Transformer-based methods, i.e., MedT [15], MCTrans [30], Swin-UNet [44], and SegFormer [45] are regarded as important methods for comparison.

2) **Implementation Details**: To ensure a fair comparison with previous studies, we resize the input images from Clean-CC-CII, JSRT, Montgomery, and NIH datasets to 512 × 512 during both training and testing phases. For the ISIC-2018 and Bowl datasets, we resize the images to 256 × 256. All input images from the GlaS dataset are resized to 128 × 128 uniformly. The TSA module in the proposed TransAttUnet framework is designed to use 8 parallel attention heads. We use stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0001 to optimize the training process. The implementation of the proposed TransAttUnet model is done using PyTorch deep learning toolbox, and all experiments are run on a single Nvidia Titan XP GPU. We train the TransAttUnet framework for 100 epochs with a batch size of 4, and the initial learning rate is set to 0.0001, which decays by a factor of 10 every 40 epochs. During our experiments, we report the probability maps generated directly by the models as the results. The probability maps are binarized using a threshold of 0.5 to obtain binary masks for performance evaluation.

3) **Evaluation Metrics**: In our experimental setup, we employ the mean Dice coefficient (DICE) as the primary evaluation metric for assessing the similarity between the predicted mask and the ground truth. Additionally, we compute four other evaluation metrics, namely mean Intersection over Union (IoU), accuracy (ACC), recall (REC), and precision (PRE) scores, on a per-pixel basis to quantify the segmentation performance. These metrics are computed based on the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values,

\[
\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN},
\]

\[
\text{IoU} = \frac{TP}{TP + FP + FN},
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},
\]

\[
\text{Recall} = \frac{TP}{TP + FN},
\]

\[
\text{Precision} = \frac{TP}{TP + FP},
\]

(C10)

3) **Multi-scale context approaches**: Meanwhile, several multi-scale context models are used as the major contenders, including Unet++ [9], R2U-Net [38] ResUNet [41], ResUNet++ [42], BCDU-Net [43], KiU-Net [37], and DoubleU-Net [36].

4) **Transformer-based approaches**: Moreover, various Transformer-based methods, i.e., MedT [15], MCTrans [30], Swin-UNet [44], and SegFormer [45] are regarded as important methods for comparison.

2) **Implementation Details**: To ensure a fair comparison with previous studies, we resize the input images from Clean-CC-CII, JSRT, Montgomery, and NIH datasets to 512 × 512 during both training and testing phases. For the ISIC-2018 and Bowl datasets, we resize the images to 256 × 256. All input images from the GlaS dataset are resized to 128 × 128 uniformly. The TSA module in the proposed TransAttUnet framework is designed to use 8 parallel attention heads. We use stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0001 to optimize the training process. The implementation of the proposed TransAttUnet model is done using PyTorch deep learning toolbox, and all experiments are run on a single Nvidia Titan XP GPU. We train the TransAttUnet framework for 100 epochs with a batch size of 4, and the initial learning rate is set to 0.0001, which decays by a factor of 10 every 40 epochs. During our experiments, we report the probability maps generated directly by the models as the results. The probability maps are binarized using a threshold of 0.5 to obtain binary masks for performance evaluation.

3) **Evaluation Metrics**: In our experimental setup, we employ the mean Dice coefficient (DICE) as the primary evaluation metric for assessing the similarity between the predicted mask and the ground truth. Additionally, we compute four other evaluation metrics, namely mean Intersection over Union (IoU), accuracy (ACC), recall (REC), and precision (PRE) scores, on a per-pixel basis to quantify the segmentation performance. These metrics are computed based on the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values,
Fig. 4. Quantitative results of the proposed TransAttUnet and state-of-the-art baselines are compared on three datasets: (a) ISIC 2018 dataset, (b) combination of JSRT, Montgomery, and NIH datasets, (c) Clean-CC-CCII dataset.

our TransAttUnet achieves the highest scores on nearly all evaluation metrics and produces a more precise and refined segmentation output than existing baselines, as illustrated in Fig. 4(a).

5) Both the diagnostic sensitivity of the proposed TransAttUnet_R and TransAttUnet_D outperform TransAttUnet_C, which demonstrates that the residual or dense step-growth connections can better aggregate multiple decoder features of varying semantic scales than the one-off cascade connections. Furthermore, the proposed TransAttUnet_R surpasses the DICE score of TransAttUnet_D by 0.6%. That is probably because the dense connection used in TransAttUnet might lead to a fair amount of redundant information. Therefore, the comparative results presented in the experiments indicate that the proposed TransAttUnet outperforms the existing state-of-the-art models on the task of automated skin lesion segmentation.

2) Evaluation on Lung Field Segmentation: We further assess the performance of the proposed TransAttUnet framework on the task of lung field segmentation using a combined dataset of JSRT, Montgomery, and NIH chest X-ray images. All evaluation results and corresponding quantitative results are shown in Table III and Fig. 4(b), respectively.
According to the findings presented in Table III, we can draw some new insights as follows:

1) Firstly, it is apparent that the evaluation metrics' corresponding scores are quite similar. This is mainly because the testing chest X-ray images provided by the JSRT and Montgomery datasets are usually normal, and the noise interference caused by various lesion features is reduced during semantic segmentation.

2) However, TransAttUnet outperforms previous state-of-the-art models and achieves the highest DICE score of 98.88% for lung field segmentation, as shown in Fig. 4(b). The segmentation outputs of TransAttUnet are closer to the ground truths compared to other baselines.

3) TransAttUnet performs significantly better than previous baselines, especially U-Net, with an improvement of 2.71% in terms of the DICE score. This improvement can be attributed to the encoder-decoder guided attention and multi-scale skip connections, which help to learn global contextual information and discriminative features to distinguish the lung field from the surrounding structures.

4) Likewise, the proposed TransAttUnet consistently outperforms the recent works, i.e., FANet (98.28%) and PraNet (98.36%), which can verify the abilities of TransAttUnet in improving the segmentation quality of the details.

3) Evaluation on Pneumonia Lesion Segmentation: Recently, pneumonia lesion segmentation has become a significant topic in the field of medical image analysis. Identifying and segmenting pneumonia lesions accurately and completely from chest CT images is challenging due to their complex textures and shapes. To further evaluate the effectiveness of the proposed TransAttUnet in this regard, we conducted experiments on the Clean-CC-CCII dataset. The comparative evaluation results are provided in Table IV, and Fig. 4(c) depicts the corresponding quantitative results.

Based on the results presented above, the following observations can be made:

1) The proposed TransAttUnet consistently outperforms the previous baselines and achieves the highest DICE score of 86.57%. These findings further confirm the superiority of TransAttUnet in medical image segmentation.

2) Both TransAttUnet_D and TransAttUnet_R are able to obtain better segmentation performance than TransAttUnet_C and the previous baselines. By contrast, the proposed TransAttUnet_R achieves a better performance than TransAttUnet_D (86.57% vs. 86.08%), which further verifies the superiorities of the residual connection in comparison with the dense connection.

3) Despite the superior performance of TransAttUnet, some reliable models, such as PraNet, obtain relatively low DICE scores in COVID-19 pneumonia lesion segmentation. This could be due to the overfitting problem resulting from the lack of training data.

4) Besides, the proposed TransAttUnet outperforms recent Transformer-based methods, such as Swin-Unet (84.47%) and SegFormer (84.96%), demonstrating the powerful ability of TransAttUnet in medical image segmentation.

5) Furthermore, as shown in Fig. 4(c), our TransAttUnet performs better than the other methods in lesion segmentation at different scales, further confirming the effectiveness of the proposed method in COVID-19 pneumonia lesion segmentation.

4) Evaluation on Nuclei Segmentation: In this part, we conducted an evaluation of the proposed TransAttUnet on the 2018 Data Science Bowl dataset for multiple nuclei segmentation. The results of this evaluation are presented in Table VI, and Fig. 5(a) provides a visualization of the corresponding quantitative results.

After analyzing the comparative results, we have made the following observations regarding the performance of TransAttUnet:

1) The use of encoder-decoder guided attention and multi-scale skip connections has improved the segmentation
quality compared to the vanilla U-Net, with the DICE score improving from 75.76% to 91.92%.
2) Consistently, the proposed TransAttUnet is superior to the existing competitors of the attention-guided and multi-scale context approaches, such as Attention U-Net (91.62% vs. 90.93%) and ResUNet (91.62% vs. 89.91%).
3) It can be seen that our TransAttUnet yields the highest score on almost all evaluation metrics. Although DoubleU-Net outperforms in terms of precision, our TransAttUnet produces the higher scores on other metrics,
especially for the IoU score of 84.98% with the improvement of 0.91%.

4) The visual comparison in Fig. 5(a) shows that TransAttUnet is effective in capturing the boundaries of cell nuclei and generating accurate segmentation predictions for multiple objects. These findings provide strong evidence for the efficacy of the proposed TransAttUnet method for multiple nuclei segmentation.

5) Evaluation on Gland Segmentation: Furthermore, we also conduct comparative experiments on the GLAS dataset to demonstrate the validity of the proposed TransAttUnet for quantifying the morphology of glands. The comparison results of evaluation metrics are presented in Table V and the corresponding quantitative results are illustrated in Fig. 5(b).

Based on the results of our experiments, we have the following observations:

1) Although the GLAS dataset is small and contains multiple complex objects, our TransAttUnet, which uses encoder-decoder guided attention and multi-scale skip connections, achieves better performance than existing baselines. This approach represents a new state-of-the-art technique for automatic gland segmentation.

2) In particular, the proposed TransAttUnet outperforms the recent Transformer-based work, i.e., MedT (81.82%), Swin-Unet (86.70%) and SegFormer (87.36%), which again demonstrates the powerful ability of TransAttUnet in medical image segmentation.

3) Furthermore, TransAttUnet_D and TransAttUnet_R outperform TransAttUnet_C and the previous state-of-the-art method, KiU-Net. In particular, TransAttUnet_D improves the DICE and IoU scores by 5.86% and 8.85%, respectively, which proves the reliability and superiority of our approach.

4) Besides, our TransAttUnet can better distinguish the gland from surrounding tissue, as shown in Fig. 5(b), resulting in excellent gland segmentation performance.

In addition, the convergence for the training loss and DICE score of our proposed TransAttUnet on different datasets is illustrated in Fig. 6. All results in the above experiments quantitatively demonstrate the effectiveness and generalizability of the proposed TransAttUnet for medical image segmentation in various challenging scenarios.

D. Ablation Studies

1) Ablation Study (I): To assess the impact of each added component in TransAttUnet, we perform a series of ablation experiments by systematically removing the components.

Table VII summarizes the experimental results. Specifically, the components are represented by the abbreviations “TSA” for transformer self-attention block, “GSA” for global spatial attention block, “MSC” for multi-scale skip connections, and “TAU” for the complete model.

We observe that when both TSA and GSA blocks are removed, the performance of “TAU w/o TSA + GSA” significantly degrade. However, it consistently outperforms the vanilla U-Net, indicating the effectiveness of the multi-scale skip connections between the decoders. Furthermore, the results of “TAU w/o TSA + MSC” and “TAU w/o GSA + MSC” demonstrate that both TSA and GSA blocks are equally effective in improving segmentation quality. We also conduct additional ablation studies by removing multiple components to further verify the effectiveness of our work. When we remove the MSC component, the evaluation scores of “TAU w/o MSC” drop significantly. Similarly, the performance of “TAU w/o TSA” and “TAU w/o GSA” degrade when we remove the other two components in turn. Overall, the experimental results prove the superiority of the designed components, and all components complement and reinforce each other to improve segmentation performance.

2) Ablation Study (II): To further investigate the impact of the SAA module’s location within the U-shaped architecture, we also conduct the ablation study with different sets on five datasets, and the comparison results are shown in Table VIII.

Specifically, our proposed TransAttUnet framework consists of a contracting path and an expansive path, where the contracting path and expansive path contain four max-pooling layers and upsampling layers, respectively. Thus, we attempt to place the SAA module in different locations and finally form six variants summarized as follows: TransAttUnet-I represents the SAA module placed between the first and second max-pooling layers;
Fig. 7. Visualizations of feature maps generated by vanilla U-Net and the proposed TransAttUnet at different decoder stages based on the Clean-CC-CCII dataset. Best viewed with zoom in.

### TABLE VIII

| Methods       | Skin  | Lung  | Pnu.  | Bowl  | Gland  |
|---------------|-------|-------|-------|-------|--------|
| TransAttUnet-I| 87.83 | 96.74 | 84.40 | 87.88 | 84.06  |
| TransAttUnet-II| 88.22 | 97.54 | 85.15 | 89.43 | 86.67  |
| TransAttUnet-III| 89.93 | 98.16 | 85.50 | 90.65 | 88.19  |
| TransAttUnet-IV| 89.30 | 98.08 | 85.32 | 90.89 | 88.24  |
| TransAttUnet-V | 88.91 | 97.98 | 84.43 | 88.54 | 87.13  |
| TransAttUnet-VI| 87.14 | 97.51 | 83.21 | 86.65 | 84.38  |
| TransAttUnet | 90.74 | 98.88 | 86.57 | 91.62 | 89.11  |

TransAttUnet-II denotes the SAA module placed between the second and third max-pooling layers; TransAttUnet-III indicates the SAA module placed between the third and fourth max-pooling layers; Similarly, TransAttUnet-IV, TransAttUnet-V and TransAttUnet-VI represent different locations of the SAA module among upsampling layers, respectively; and TransAttUnet denotes the way that the SAA module is located at the bottom of the U-shaped architecture.

From Table VIII, we can observe that any degenerated variants are inferior to our TransAttUnet model. Actually, the SAA module can be placed anywhere in the network as a plug-and-play module. However, by locating it at the bottom of the architecture, the SAA module can effectively capture the rich spatial information of the input data and provide a stronger foundation for subsequent feature extraction and aggregation.

### E. Visualizations of Decoder Stages

To demonstrate the superior performance of TransAttUnet over the vanilla U-Net, we conducted a visualization analysis of the feature maps of each decoder stage. This analysis aimed to highlight the TransAttUnet’s ability to capture long-range feature dependencies and global contextual information, which enhances its semantic discrimination capabilities. Fig. 7 shows the visualizations of the feature maps of each decoder stage for both U-Net and TransAttUnet.

Based on the comparative results, we can make the following observations:

1) It can be seen that the encoders of U-Net fail to make full use of the contextual information. With the guidance of the multi-level non-local attention mechanisms that effectively capture contexts, the obtained encoder features can provide more global semantic information to our decoders at lower stages, which can generate the most discriminative features.

2) Meanwhile, our TransAttUnet can make full use of the multi-scale contextual information to generate accurate predictions at later stages. As shown with the zoom-in patches of the deepest stage, the segmentation results of TransAttUnet are more detailed and reliable with a clearly discernible boundary. Hence, we can infer that our approach learns more effective semantic information, resulting in better performance for medical image segmentation.

### V. CONCLUSION AND FUTURE WORK

In this article, we propose a novel Transformer based attention-guided U-Net called TransAttUnet, which concurrently incorporates multi-level guided attention and multi-scale skip connections into U-Net for improving the segmentation quality of biomedical images. Specifically, the multi-level guided attention block is able to make full use of global contextual information by concurrently exploring long-range interactions and global spatial relationships among encoder semantic features. Meanwhile, the multi-scale skip connection scheme can flexibly aggregate contextual feature maps from decoders of varying semantic scales to generate the discriminative feature representations. Compared to previous state-of-the-art methods, our proposed TransAttUnet model shows significant improvements in medical image segmentation by utilizing long-range feature dependencies and multiscale contextual information to create feature representations with greater semantic consistency. In this way, we can effectively mitigate the intrinsic limitations...
that occur in traditional U-shape architecture. Extensive experimental results on different benchmark datasets demonstrate that the proposed TransAttUnet can achieve consistent performance improvements by integrating the above novelties. Despite the fact that our study has some vital contributions, there are still several limitations. Admittedly, the proposed TransAttUnet heavily relies on the global self-attention mechanism and thus suffers the large memory footprint and computation cost. In addition, the potential of Transformer for medical image segmentation remains incomplete and underutilized in the study of our work, especially in the face of various biomedical images. Moreover, we will attempt to introduce the big model as a fundamental model for segmenting medical images. Therefore, further improvements in these aspects will be investigated in our future work.

REFERENCES

[1] S.-G. Lee, E. Kim, J. S. Bae, J. H. Kim, and S. Yoon, “Robust End-to-End focal liver lesion detection using unregistered multiphase computed tomography images,” IEEE Trans. Emerg. Topics Comput. Intell., vol. 7, no. 2, pp. 319–329, Apr. 2023.

[2] X. Yao, Z. Zhu, C. Kang, S.-H. Wang, J. M. Gorriz, and Y.-D. Zhang, “AdaD-FNN for chest CT-based COVID-19 diagnosis,” IEEE Trans. Emerg. Topics Comput. Intell., vol. 7, pp. 5–14, Feb. 2023.

[3] H. Liu, M. J. Grothe, T. Rashid, M. A. Labrador-Espinosa, J. B. Toledo, and M. Habes, “ADCoC: Adaptive distribution modeling based collaborative clustering for disentangling disease heterogeneity from neuroimaging data,” IEEE Trans. Emerg. Topics Comput. Intell., vol. 7, no. 2, pp. 308–318, Apr. 2022.

[4] J. Fu et al., “Contextual deconvolution network for semantic segmentation,” Pattern Recognit., vol. 101, 2020, Art. no. 107152.

[5] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. 15th Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2015, pp. 234–241.

[7] W. Luo, Y. Li, R. Urtasun, and R. Zemel, “Understanding the effective receptive field in deep convolutional neural networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2016, pp. 4905–4913.

[8] Y. Chen et al., “Channel-Unet: A spatial channel-wise convolutional neural network for liver and tumors segmentation,” Front. Genet., vol. 10, 2019, Art. no. 1110.

[9] Z. Zhou, M. M. R. Siddique, N. Tajbakhsh, and J. Liang, “Unet: A nested U-Net architecture for medical image segmentation,” in Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support, 2018, pp. 3–11.

[10] O. Oktay et al., “Attention U-net: Learning where to look for the pancreas,” Med. Image Anal., vol. 53, no. 2, pp. 197–207, 2019, doi: 10.1016/j.media.2019.01.012.

[11] Y. Cai and Y. Wang, “Pancreas,” IEEE Trans. Biomed. Eng., vol. 66, no. 5, pp. 989–1001, May 2018.

[12] P. Malplík, S. Kristofik, and K. Knapová, “Instance segmentation model created from three semantic segmentations of mask, boundary and centroid pixels verified on Gluas dataset,” in Proc. IEEE 15th Conf. Comput. Sci. Inf. Syst., 2020, pp. 569–576.

[13] P. Zhao, J. Zhang, W. Fang, and S. Deng, “SCAU-Net: Spatial-channel attention Unet for gland segmentation,” Front. Bioeng. Biotechnol., vol. 8, 2020, Art. no. 670.

[14] M. Islam, V. Vishwan, V. J. M. Jose, N. Wijeathilake, U. Utkarsh, and H. Ren, “Brain tumor segmentation and survival prediction using 3D attention Unet,” in Proc. Int. MICCAI Brainlesion Workshop, 2019, pp. 262–272.

[15] O. Čípek, A. Abdulkašir, S. L. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in Proc. 19th Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2016, pp. 424–432.

[16] Y.-B. Tang, Y.-X. Tang, J. Xiao, and R. M. Summers, “XLSor: A robust and accurate lung segmentor on chest X-rays using criss-cross attention and customized radiological abnormalities generation,” in Proc. Int. Conf. Med. Imag. Deep Learn., 2019, pp. 457–467.

[17] M. Salehi, M. A. Ardekani, A. B. Taramsari, H. Ghaffari, and M. Haghparast, “Automated deep learning-based segmentation of COVID-19 lesions from chest computed tomography images,” Pol J Radiol., vol. 87, pp. 478–482, 2020.

[18] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jager, “U2-Net: Going deeper with nested u-structure for salient object detection,” Pattern Recognit., vol. 106, 2020, Art. no. 107404.

[19] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2020.

[20] 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.

[21] S. Zhang et al., “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 6881–6890.

[22] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” in Proc. 24th Med. Image Comput. Comput. Assist. Interv., 2021, pp. 14–24.

[23] Y. Li et al., “Multi-compound transformer for accurate biomedical image segmentation,” in Proc. 24th Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2021, pp. 326–336.

[24] N. Codella et al., “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC),” in Proc. Int. Symp. Biomed. Imag., 2019, pp. 168–172.

[25] J. Shiroyashii et al., “Development of a digital image database for chest radiographs with and without lung nodules: Receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules,” Amer. J. Roentgenol., vol. 174, no. 1, pp. 71–74, 2000.

[26] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, and G. Thoma, “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases,” Quantitative Imag. Med. Surg., vol. 4, no. 6, 2014, Art. no. 475.

[27] X. He et al., “Automated model design and benchmarking of deep learning models for covid-19 detection with chest ct scans,” in Proc. AAAI Conf. Artif. Intell., 2021, vol. 35, no. 6, 2021, pp. 4821–4829.

[28] J. C. Caicedo et al., “Nucleus segmentation across multiple imaging environments: The 2018 data science bowl,” Nature Methods, vol. 16, no. 12, pp. 1247–1253, 2019.

[29] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, “Doubleu-net: A deep convolutional neural network for medical image segmentation,” in Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst., 2020, pp. 558–564.

[30] J. M. J. Valanarasu, V. A. Sindagi, I. Hachihalioglu, and V. M. Patel, “Kner-net: Towards accurate segmentation of biomedical images using over-complete representations,” in Proc. 23rd Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2020, pp. 363–373.

[31] M. Z. Alom, C. Yakopcic, T. M. Taha, and V. K. Asari, “Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net),” in Proc. IEEE Nat. Aerosp. Electron. Conf., 2018, pp. 226–233.

[32] N. K. Tomar et al., “FANet: A feedback attention network for improved biomedical image segmentation,” IEEE Trans. Neural Netw. Learn. Syst., pp. 1–14, 2022.
[40] D.-P. Fan et al., “PraNet: Parallel reverse attention network for polyp segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 263–273.

[41] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, “Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.

[42] D. Jha et al., “ResUNet: An advanced architecture for medical image segmentation,” in *Proc. IEEE Int. Symp. Multimedia*, 2019, pp. 225–2255.

[43] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, “Bi-directional ConvLSTM U-Net with densely connected convolutions,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1–10.

[44] H. Cao et al., “Swin-UNet: Unet-like pure transformer for medical image segmentation,” in *Proc. Comput. Vis.–ECCV, Part III*, 2022, pp. 205–218.

[45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

**Bingzhi Chen** received the Ph.D. degree in computer applied technology from the Harbin Institute of Technology, Harbin, China. He is currently with the School of Software, South China Normal University, Foshan, Guangdong, China, and also with the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies. His research interests include medical diagnosis, pattern recognition, and deep learning.

**Yishu Liu** received the B.S. and M.S. degrees from the Harbin Institute of Technology, Harbin, China, in 2019 and 2022, respectively. She is currently working toward the Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. Her research interests include computerized medical diagnosis, pattern recognition, deep learning, and machine learning.

**Zheng Zhang** (Senior Member, IEEE) received the Ph.D. degree in computer applied technology from the Harbin Institute of Technology, Shenzhen, China. He was a Postdoctoral Research Fellow with The University of Queensland, Brisbane, QLD, Australia. He is currently with the Harbin Institute of Technology, Shenzhen, China. He has authored or coauthored more than 100 technical papers in prestigious international journals and conferences. His research interests include multimedia content analysis and understanding. He is an Associate Editor for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS.

**Guangming Lu** (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the Harbin Institute of Technology (HIT), Harbin, China, and 2005, respectively. He is currently a Professor with the Harbin Institute of Technology, Shenzhen, China, Associate Dean of the School of Computer Science, Harbin Institute of Technology (Shenzhen), Deputy Director of the Guangdong Provincial Key Laboratory of Security Intelligent New Technology, and Director of the Shenzhen Medical Biometric Perception and Analysis Technology Engineering Laboratory. He has authored or coauthored more than 200 technical papers at prestigious international journals and conferences. His research interests include pattern recognition, image processing, and automated biometric technologies and applications.

**Adams Wai Kin Kong** (Member, IEEE) received the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada. He is currently an Associate Professor with the Nanyang Technological University, Singapore, and the Director of Master of Science in Artificial Intelligence programme. His research interests include pattern recognition, deep learning, and their applications on power systems, healthcare, and biometrics.