Computing and Informatics, Vol. 42, 2023, 392–410, doi: 10.31577/cai_2023_2_392

CTRANSNET: CONVOLUTIONAL NEURAL NETWORK COMBINED WITH TRANSFORMER FOR MEDICAL IMAGE SEGMENTATION

Zhixin Zhang, Shuhao Jiang, Xuhua Pan*

Information Engineering Department
Tianjin University of Commerce
Tianjin, 300134, China
e-mail: zhangzhixin010101@163.com, {Jiangshuhao, Panxuha}@tjcu.edu.cn

Abstract. The Transformer has been widely used for many tasks in NLP before, but there is still much room to explore the application of the Transformer to the image domain. In this paper, we propose a simple and efficient hybrid Transformer framework, CTransNet, which combines self-attention and CNN to improve medical image segmentation performance. Capturing long-range dependencies at different scales. To this end, this paper proposes an effective self-attention mechanism incorporating relative position information encoding, which can reduce the time complexity of self-attention from $O(n^2)$ to $O(n)$, and a new self-attention decoder that can recover fine-grained features in encoder from skip connection. This paper aims to address the current dilemma of Transformer applications: i.e., the need to learn induction bias from large amounts of training data. The hybrid layer in CTransNet allows the Transformer to be initialized as a CNN without pre-training. We have evaluated the performance of CTransNet on several medical segmentation datasets. CTransNet shows superior segmentation performance, robustness, and great promise for generalization to other medical image segmentation tasks.

Keywords: Medical image segmentation, deep learning, attention mechanism

* Corresponding author

1 INTRODUCTION

With the development and widespread use of medical imaging equipment, X-rays, CT examinations, magnetic resonance imaging (MRI), and ultrasound scans have become four necessary medical aids used to assist doctors in disease diagnosis, prognosis assessment, and surgery planning. To help doctors make accurate diagnoses, medical image segmentation is required to identify some critical targets in medical images and extract features from them for subsequent lesion diagnosis. In general, there are two main types of image segmentation tasks: semantic segmentation and instance segmentation. Image semantic segmentation is a pixel-level classification task that requires predicting each pixel point of an image. Image instance segmentation requires not only pixel-level classification but also the differentiation of instances based on specific categories. Medical image segmentation is unique in that there are significant differences between each organ or tissue, making instance segmentation of medical images less meaningful. Medical image segmentation usually refers to the semantic segmentation of medical images. Currently, the main medical image segmentation tasks include liver and liver tumour segmentation, brain and brain tumour segmentation, optic disc segmentation, cell segmentation, lung segmentation, and lung nodule segmentation. Many recent medical semantic segmentation approaches have adopted the U-Net framework with a codec structure. However, U-Net using a simple jump-join scheme is still challenging for modelling global multi-scale problems:

1. Not every jump-join setting is valid due to incompatible codec stage feature sets, and even some jump-join can negatively affect segmentation performance;
2. The original U-Net is worse than U-Net without jump-join on some datasets.

CNNs are widely used in computer vision tasks because of their excellent feature extraction capabilities; the encoder-decoder structure built on convolutional operations is currently well-suited for solving location-sensitive tasks such as semantic segmentation. With the help of convolution operations, texture information and local features between neighbouring pixels can be captured; then, by stacking the local features extracted at different levels, the perceptual field can be gradually expanded to obtain higher-level global features. However, this approach has two limitations: firstly, convolution can only extract information between neighbouring pixels and cannot model global associations effectively; secondly, the size and dimensions of the convolution kernel are often fixed and cannot be adjusted according to the input content.

The Transformer has been widely used for many tasks in NLP before [1, 2, 3], but there is still much room to explore the application of the Transformer to the image domain [4, 5, 6]. In this paper, we propose a simple and efficient hybrid Transformer framework, CTransNet, which combines self-attention and CNN to improve medical image segmentation performance and capturing long-range dependencies at different scales. To this end, this paper proposes an effective self-attention mechanism incorporating relative position information encoding, which can reduce the

time complexity of self-attention from $O(n^2)$ to $O(n)$, and a new self-attention decoder that can recover fine-grained features in encoder from skip connection. This paper aims to address the current dilemma of Transformer applications: i.e., the need to learn induction bias from large amounts of training data. The hybrid layer in CTransNet allows the Transformer to be initialized as a CNN without pre-training. We have evaluated the performance of CTransNet on several medical segmentation datasets. CTransNet shows superior segmentation performance, robustness, and great promise for generalization to other medical image segmentation tasks.

Based on the above findings, we propose a new medical image segmentation framework, CTransNet, which leverages channel attention mechanisms. Our approach utilizes a hierarchical cascaded self-attention module (MHSA) to address the inefficiency of multi-headed self-attention in the visual Transformer model caused by high computational and spatial complexity. We propose to split the image into patches, with each patch representing a token to learn feature relationships within a small grid. We group patches into each small grid and compute self-attention in each group, capturing local feature relationships and producing different local feature representations. The smaller grids are then merged into the larger grid, with the previous smaller grid treated as a new token for the next grid's attention computation. CTransNet combines self-attention [7] and convolutional neural network (CNN) techniques to improve segmentation performance, with self-attention modules incorporated into both the encoder and decoder parts to capture long-range dependencies at different scales with minimal overhead. Our approach uses an effective self-attention mechanism that includes relative position information encoding to reduce self-attention's time complexity from $O(n^2)$ to $O(n)$. Additionally, our self-attention decoder can recover fine-grained features in the encoder from skip connection. Experimental results demonstrate that CTransNet outperforms traditional architectures, including transformer and U-Shape frameworks, across different datasets, leading to more accurate and consistent improvements in semantic segmentation.

2 RELATED WORK

2.1 CNN-Based Methods

Early methods for segmenting medical images primarily relied on contour and conventional machine learning techniques [8, 9, 2, 10]. U-Net for medical picture segmentation was proposed in [11] with the introduction of deep CNNs. Numerous Unet-like techniques, like Res-UNet [12], have been developed as a result of the U-shaped structure's ease of use and high performance. U-Net++ [13], Dense-UNet [10], and UNet3+ [14]. Additionally, it has been applied to the segmentation of 3D medical images using methods like 3D-Unet [15] and V-Net [16]. In the field of medical picture segmentation right now, CNN-based techniques have had remarkable success. Because of its potent representation, CNN-based tech-

niques have now had tremendous success in the field of medical image segmentation.

2.2 Vision Transformers Methods

Transformer was initially put forth as a solution for machine translation tasks in [17]. In the field of NLP, techniques based on transformers have excelled in a range of tasks, achieving state-of-the-art performance. Multiple tasks have been completed with state-of-the-art performance [18]. Due to the popularity of Transformer, researchers at [19] developed the ground-breaking Visual Transformer (ViT), which demonstrated a remarkable speed-accuracy trade-off in picture recognition tasks. Because ViT needs pre-training on its own sizable dataset, it is less advantageous than CNN-based techniques. Deit et al. [20] outlines numerous training procedures that make it possible for ViT to be effectively trained on ImageNet in order to overcome the challenges associated with doing so. Recently, some outstanding papers on ViT have been produced [21, 22, 23]. Notably, the Swin Transformer was proposed as the visual backbone given in [23], and it is an effective hierarchical visual transformer. The Swin Transformer, which is based on the shift window mechanism, performs at the cutting edge on a range of vision tasks, including semantic segmentation, object detection, and image classification. In this paper, we try by employing the Swin Transformer block as the basis unit to create a U-shaped encoder-decoder architecture with skip connections for medical picture segmentation, we want to provide a benchmark for the advancement of transformers in the field of medical images. A benchmark comparison can be made using the Transformer's advancement in the realm of medical pictures.

2.3 Transformer to Complement CNNs

In recent years, researchers have attempted to increase network performance by incorporating self-attention mechanisms into CNNs [24, 25, 26, 27]. There are also a number of vision tasks on which Transformer and CNN [28, 29] have been combined, and significant improvements have been achieved. In [30], a U-shaped structure was integrated with skip connections and additive attention gates to analyse medical images. However, this strategy is still CNN-based. Efforts are currently being made to combine CNN with Transformer to challenge CNNs dominance in medical image segmentation. CNNs have advantages for medical picture segmentation [25, 31, 32]. The authors of [25, 33] have developed a potent encoder for the segmentation of two-dimensional medical images. Similar to [25, 31] and [34] use the complementary nature of the Transformer and CNN to enhance the segmentation capabilities of the model. Various combinations of Transformer and CNN are currently employed for the multimodal segmentation of brain tumors [35] and 3D medical picture segmentation [32, 2]. In contrast to the methodologies described above, we investigated the possibility of pure transformers for medical image segmentation applications. We redesigned the multi-headed attention mechanism of

the Transformer and perfectly fused the local information extraction of the convolutional neural network and the global context of the Transformer to make our method more applicable to image segmentation tasks.

3 METHODOLOGY

3.1 Self-Attention

The Transformer model is founded on a multi-head attention module (MHSA, Multi-head self-attention) that enables the model to incorporate attention learned from different subspaces. The output of the multi-heads is concatenated and supplied to the feedforward network (FFN) layer. Given the small sample size of medical datasets, we conducted several experiments and found that a large number of parameter calculations could have an adverse impact on model segmentation performance. Therefore, we determined that the head parameter setting of 6 achieved the best performance for our method. In this study, we applied head = 6 to the input X ($C \times W \times H$) to obtain the mapping Q, K, V after a 1×1 convolution, which is then divided into various heads. The following equation outlines the specific attention calculation:

$$Att(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{1}$$

The computed attention is then processed by softmax and called: a contextual aggregation matrix, or similarity matrix, indicating how well each q matches is similar to all the keys; this similarity is then used as a weight and multiplied by the value, so that attention is computed, and is based on a global perceptual field that takes all the inputs into account. This self-attention-based contextual aggregation matrix dynamically adjusts with the input content, allowing for better feature aggregation; however, the dot product operation for $n \times d$ has a time complexity of $O(n^2)$, as n as a sequence length is generally much larger than the dimension d. Where $\sqrt{d}$ denotes approximate normalization, applying the Softmax function to each row of the matrix. Note that we have omitted the computation of multiple headers here for simplicity. The matrix product QK^T is done specifically by first computing the similarity between each pair of tokens. Then, each new token is obtained by derivative acquisition on top of the combination of all tokens. after the MHSA calculation, further residual joins can be added to facilitate optimization. We assume that the height of the input X feature map is H_0 and the width is W_0. We have $N = H_0 \times W_0$. Then, we can divide the feature map into small grids, each of size $G_0 \times G_0$. Therefore, we reconstruct the input feature map to obtain the new X':

$$X \in R^{C\times H_0\times W_0} \to X \in R^{C\times\left(\frac{H_0}{G_0}\times G_0\right)\times\left(\frac{W}{G_0}\times G_0\right)} \to X' \in R^{C\times\left(\frac{H_0}{G_0}\right)\times\left(\frac{W_0}{G_0}\right)\times(G_0\times G_0)}. \tag{2}$$

To simplify the network optimization, we also perform the following transformation for the generated local self-attentive Att:

$$Att_0 \in R^{C\times H_0\times W_0} \rightarrow Att_0 \in R^{C\times\left(\frac{H_0}{G_0}\times G_0\right)\times\left(\frac{W}{G_0}\times G_0\right)} \rightarrow Att_0' \in R^{C\times\left(\frac{H_0}{G_0}\right)\times\left(\frac{W_0}{G_0}\right)\times(G_0\times G_0)}. \tag{3}$$

This computational complexity is significantly reduced because the Att_0 computes each small $G \times G$ network faster. For the i^{th} step, we can consider the smaller network block obtained at the $i-1^{\text{st}}$ step as a new token, which can be achieved simply by downsampling the attentional features:

$$Att_0 = X + Att_0, \tag{4}$$

$$Att_{i-1}' = MaxPool(Att_{i-1}) + AvgPool(Att_{i-1}), \tag{5}$$

where $Att_{i-1}' \in R^{C\times H_i\times W_i}$, $H_i = H_0/(G_0G1\ldots G_{i-1})$, $W_i = W_0/(G_0G1\ldots G_{i-1})$, $MaxPool$ and $AvgPool$ denote maximum pooling and average pooling, respectively. We then similarly divide Att_{i-1}' into a grid of size $G_i \times G_i$ and re-obtain the following equation:

$$\begin{aligned} Att_{i-1}' \in R^{C\times H_i\times W_i} &\rightarrow Att_{i-1}' \in R^{C\times\left(\frac{H_i}{G_i}\times G_i\right)\times\left(\frac{W}{G_i}\times G_i\right)} \\ &\rightarrow Att_{i-1}' \in R^{C\times\left(\frac{H_i}{G_i}\right)\times\left(\frac{W_i}{G_i}\right)\times(G_i^2)}, \end{aligned} \tag{6}$$

$$Q = X_{i-1}'W^q, K = X_{i-1}'W^k, V = X_{i-1}'W^v, \tag{7}$$

finally, we obtain the mathematical representation of A_i as follows:

$$Att_i' \in R^{C\times H_i\times W_i} \rightarrow Att_i' \in R^{C\times\left(\frac{H_i}{G_i}\times G_i\right)\times\left(\frac{W}{G_i}\times G_i\right)} \rightarrow Att_i' \in R^{C\times\left(\frac{H_i}{G_i}\right)\times\left(\frac{W_i}{G_i}\right)\times(G_i^2)}. \tag{8}$$

We connect through the residuals and will keep iterating until it is small enough. Then we stop slicing the grid blocks. the final output of MHSA is:

$$MHSA(X) = (Att_0 + \ldots + Upsample(Att_M))W^p + X, \tag{9}$$

where $UPsample(.)$ denotes upsampling the attentional features to their original size and W^p is the weight matrix of the feature projection. m is the maximum number of iteration steps. In this way, our method can establish global feature dependencies. It is easy to prove that, under the assumption that all G_i are equal, the computational complexity of $MHSA$ is:

$$T_{time}(MHSA) = 3NC^2 + 2NG_0^2C. \tag{10}$$

Thus, we reduce the computational complexity significantly from $O(N^2)$ to $O(N)$, and here G_0 is much smaller than N. Likewise, the space complexity is greatly reduced.

In terms of network time complexity computation, our approach differs from some state-of-the-art not-transformer-based approaches in that we first divide the image into multiple patches, each of which can be considered as a token, and instead of computing attention across all patches, we further group the patches into each small grid and compute self-attention in instead of computing attention across all patches, we further group patches into each small grid and compute self-attention in each grid, thus capturing local feature relationships and producing distinguishable local feature representations. Then, the smaller grids are merged into the larger grid, and the attention in the next grid is recomputed by treating the smaller grid in the previous step as a new token. This process is repeated iteratively to gradually reduce the number of tokens. Throughout the process, our MHSA module progressively computes self-attention in increasing regional network sizes and naturally models the global feature relationships in a hierarchical manner. Since each grid has only a small number of tokens at each step, we can significantly reduce the computational/spatial complexity of the vision Transformer.

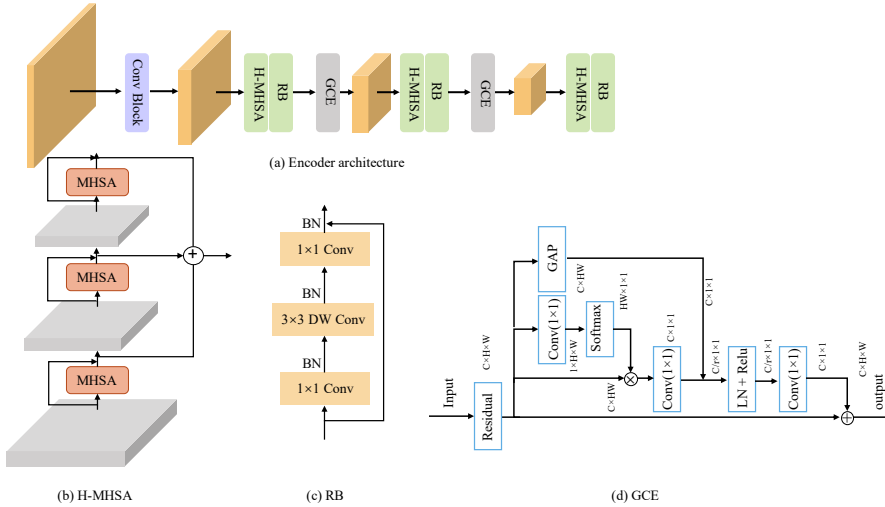

Figure 1. Illustration of the proposed CTransNet. GAP: global average pooling; DW Conv: deepthwise separable convolution; RB: Residual Bottleneck; GCE: global context extraction.

3.2 Network Architecture

Figure 1 illustrates the network structure of CTransNet. The purpose of this study is to combine the benefits of convolution with self-attention so that, on the one hand, convolution can be used to learn inductive bias and avoid pre-training the

Transformer on big datasets, and on the other hand, the Transformer can be used to capture global characteristics. The effective self-attention and relative location coding suggested in this research enable the Transformer to accumulate global contextual data at various scales efficiently. Since miss segmentation frequently happens at the edges of the ROI region, high-resolution contextual information is required for precise segmentation. Instead of only computing self-attention on the CNN-extract feature map, this research employs a transformer at each level of the encoder-decoder to collect long-range dependencies at various scales. However, the raw input was not processed using the Transformer, as employing the Transformer at a superficial level would be of limited benefit and raise the computing cost. One possible explanation for this is that the shallow feature map is more concerned with fine-grained textures than global information. Since the Euclidean distance possesses symmetry, the disease-centric learning strategy, in this case, can be substituted by r. Figure 3 depicts a symmetric metric learning approach centred on drugs and diseases under the explicit treatment relationship. In summary, the disease-centric metric is symmetric with the drug-centric metric, and the objective of symmetric metric learning is to push drugs or diseases that are not associated out of the ball, pull drugs or diseases that are associated or potential associations into the ball, and guarantee that the distance of known drug-disease pairs is smaller than the distance between unknown associations.

3.3 Loss Fuction

Our proposed approach employs the widely-used cross-entropy as the loss function, which serves as a metric to evaluate the degree of agreement between the predicted and ground-truth outputs. In the context of classification training, for a given sample belonging to the K^{th} class, the corresponding output node should have a value of 1 while the remaining nodes have values of 0, forming the target label. By calculating the cross-entropy loss function, we quantify the discrepancy between the predicted output and the target label, and use this difference to update the network parameters through backpropagation. The cross-entropy loss function measures the divergence between the predicted probability distribution and the true probability distribution, where lower cross-entropy implies greater similarity between the two distributions. Formally, assuming p and q as the target and predicted probability distributions, respectively, the cross-entropy loss function is defined as follows:

$$\mathcal{L}_{CE} = -\sum_{x}(p(x)\log q(x) + (1 - p(x)\log(1 - q(x))), \tag{11}$$

where $p(x)$ is the expected output and the probability distribution $q(x)$ is the actual output.

4 EXPERIMENTS AND DISCUSSION

In this section we will focus on some of the details and steps in the experimental process, and the comparative results of some of the most advanced methods and the visualisation of the experimental results on the graphs.

4.1 Datasets and Evaluation

4.1.1 Kvasir-SEG Datasets

Kvasir-SEG is an open-access collection of gastrointestinal polyp pictures and related segmentation masks that were manually annotated by a medical practitioner and subsequently validated by a seasoned gastroenterologist. The Kvasir-SEG dataset includes one thousand polyp pictures and their related ground truth from the Kvasir Dataset v2. The resolution of the photos contained in Kvasir-SEG ranges from 332×487 to $1\,920 \times 1\,072$ pixels. The photos and their respective masks are saved in two distinct folders with the same name. The image files are compressed using the JPEG format, which facilitates online viewing. The publicly available dataset is freely downloadable for research and teaching purposes. The bounding box (coordinate points) for the respective photos is saved in a JSON file. This data collection is intended to further the current best method for polyp identification.

4.1.2 DRIVE Datasets

The DRIVE database was designed to facilitate comparative research on the segmentation of blood vessels in retinal pictures. Retinal vessel segmentation and delineation of morphological attributes of retinal vessels, such as length, width, tortuosity, branching patterns, and angles, are utilized for the diagnosis, screening, treatment, and evaluation of numerous cardiovascular and ophthalmic diseases, such as diabetes, hypertension, atherosclerosis, and choroidal neovascularisation. Automated detection and analysis of blood vessels can help create screening programs for diabetic retinopathy, research the association between vascular tortuosity and hypertensive retinopathy, and aid in computer-assisted laser surgery. For temporal or multimodal image registration and retinal image mosaic synthesis, automatic retinogram generation and branch point extraction have been employed. In addition, it was discovered that the retinal vascular tree is unique to each individual and can be utilized for biometric purposes.

4.1.3 Evaluation

In Equation (12), the accuracy, sensitivity, IoU, and Dice are shown as a criterion group to completely evaluate the experimental outcomes.

$$\begin{cases} Accuracy &= \frac{TP+TN}{TP+FP+FN+TN}, \\ Sensitivity &= \frac{TP}{TP+FN}, \\ IoU &= \frac{TP}{TP+FN+FP}, \\ Dice &= \frac{2\times TP}{(TP+FN)+(TP+FP)}. \end{cases} \tag{12}$$

In this study, the performance of the predictive model is evaluated using several metrics, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These metrics represent the number of correctly predicted positive and negative samples, the number of negative samples that were incorrectly predicted as positive, and the number of positive samples that were incorrectly predicted as negative, respectively. Additionally, the sensitivity against specificity is assessed using the Area Under the ROC Curve (AUC) metric. This measure is commonly used to evaluate the performance of binary classification models, where sensitivity is the true positive rate and specificity is the true negative rate.

4.2 Implementation Details

Our CTransNet was implemented using the Pytorch deep learning framework, and we conducted a range of hyperparameter tuning experiments, such as adjusting the learning rate, batch size, weight decay rate, and resize parameters. Both training and testing were carried out on Ubuntu 18.04, using two RTX 2080Ti graphics cards with 12 GB of video memory each. The small batch stochastic gradient descent (SGD) method was employed for training, with a batch size of 8 and a learning rate of 0.0001 on the DRIVE dataset, and a batch size of 8 and a learning rate of 0.001 on the Kvasir-SEG dataset. We compared Adam optimization with SGD and found that SGD typically outperforms Adam, albeit at a slower convergence rate. Despite Adam converging faster, we prioritized performance in both time and accuracy. To validate the effectiveness of our approach, we conducted experiments on multiple datasets, as shown in the figure below, and demonstrated that our approach consistently achieved favourable results.

4.3 Experimental Results

4.3.1 Result on DRIVE Dataset

DRIVE is a dataset that permits the segmentation of retinal blood vessels. It consists of forty color retinal images, twenty of which are used for training and twenty of which are used for evaluation. Originally, the dimensions of the images were 565×584 pixels. A dataset sample of this size is insufficient for training a deep neural network. Consequently, we apply the following strategy to overcome this issue: Beginning with the provided images, random blocks are generated. The remaining photos

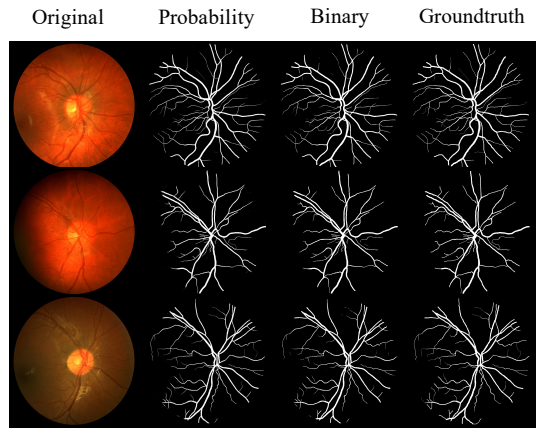

Figure 2. The segmentation results of CTransNet on DRIVE dataset

Methods	**Accuracy**	**Specificity**	**Sensitivity**	**AUC**
Backbone	0.9477	–	0.7781	0.9705
UNet [11]	0.9531	0.9820	0.7537	0.9680
R2-Uet [36]	0.9652	0.8303	0.7792	0.9245
Deep Model [37]	0.9495	0.9768	0.7763	0.9720
RU-net [38]	0.9553	0.9820	0.7726	0.9779
Attention-Unet [39]	0.9629	0.9725	0.7884	0.9740
Unet++ [13]	0.9656	0.9867	0.8234	0.9628
BCD-Unet [40]	0.9560	0.9786	0.8007	0.9789
CENet [41]	0.9545	0.9851	0.8309	0.9779
Fusion Mechanism [42]	0.8247	0.9847	0.8140	0.9782
CTtansNet(Ours)	**0.9660**	**0.9870**	**0.8433**	**0.9785**

Table 1. Performance comparison of the proposed network and the State-of-the-Art methods on DRIVE dataset

were utilized to validate 19,000 segmentation findings using DRIVE. The batch size used as input data for the network was 64×64.

The Figure 2 illustrates some precise of CTransNet and promising segmentation results. In the four columns are listed the original RGB image, the anticipated probability image, the predicted binary image, and the ground truth. Table 1 offers further state-of-the-art research and quantitative findings produced by the proposed network CTransNet on the DRIVE dataset. Our studies were assessed using five unique measures. CTransNet performs brilliantly in terms of accuracy, specificity, sensitivity, and AUC, with respective values of 0.9660, 0.9870, 0.8433, and 0.9785.

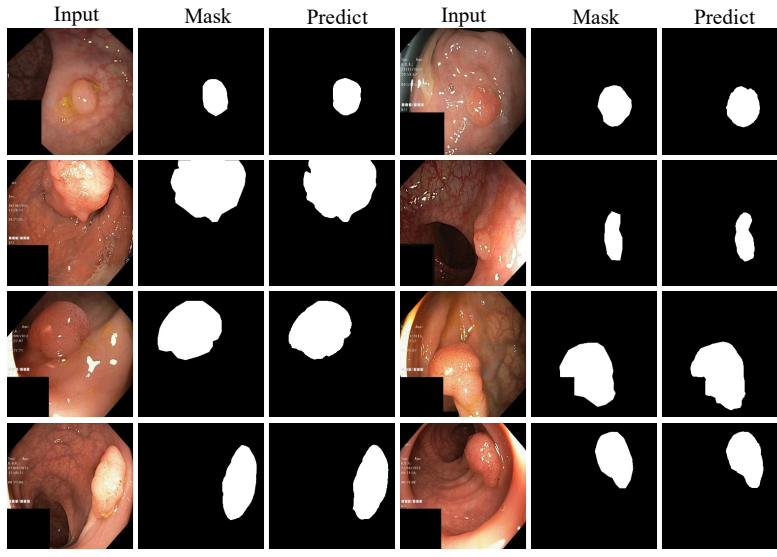

Figure 3. The segmentation results of CTransNet on Kvasir-SEG dataset datasets

4.3.2 Result on Kvasir-SEG Dataset

The results of our CTransNet visualization test on the Kvasir-SEG dataset are shown in Figure 3, from left to right, Input, Mask and Predict. It can be seen that our algorithm has a low error value with Mask. In addition, we also compared it with some classical methods, as shown in Table 2, where our method achieves state-of-the-art performance in several metrics. the values of Precision, Recall, mIOU, and Dice for UNet on the Kvasir-SEG dataset are 92.22, 63.06, 43.43, and 81.80, respectively. resUNet The values of Precision, Recall, mIOU, and Dice on the Kvasir-SEG dataset are 72.92, 50.41, 43.64, and 51.44, respectively. The values of Precision, Recall, mIOU, and Dice on the Kvasir-SEG dataset for MSRF-Net are 96.66, 91.88, 89.14, 92.17. The values of Precision, Recall, mIOU, and Dice for CTransNet on the Kvasir-SEG dataset are 96.75, 90.15, 89.32, and 93.21, respectively. MSRF-Net exceeds our Recall metric by 1.83 %, and their different perceptual fields and multi-scale residual fusion network have significant advantages for the image segmentation task. Experimental results show that our method outperforms the existing state-of-the-art methods in several evaluation metrics, and we analyze some specific reasons why our method efficiently combines visual local attention and contextual information, which is crucial for our semantic segmentation task. Experimental results show that our method outperforms existing state-of-the-art methods in several evaluation metrics, and we analyze some specific reasons why our method effectively combines visual local attention and contextual information, which is crucial for our semantic segmentation task, especially for small dataset tasks where global information is

more important. However, MSRF-Net is currently 1.83 % ahead of us in the Recall metric, which may be an advantage for the Recall metric as MSRF-Net is able to use dual-scale dense fusion to exchange multi-scale features from different perceptual fields.

Methods	**Precision**	**Recall**	**mIoU**	**Dice**
Unet [11]	92.22	63.06	43.34	81.80
ResUNet [43]	72.92	50.41	43.64	51.44
ResUnet-mod [44]	87.13	69.09	42.87	79.09
ResUnet++ [45]	70.64	70.64	79.27	81.33
DeeplabV3+ [46]	94.96	89.84	85.75	89.65
DDANet [47]	86.43	88.80	78.00	85.76
MSRF-Net [48]	96.66	**91.98**	89.14	92.17
CTransNet (Ours)	**96.75**	90.15	**89.32**	**93.21**

Table 2. Performance comparison of the proposed network and the State-of-the-Art methods on n Kvasir-SEG dataset

5 SENSITIVITY ANALYSIS

In this section, in order to verify the effective performance of our method, we conducted a series of ablation experiments aimed at verifying the role of each component on the whole network, and we chose the dataset Kvasir-SEG for this purpose, and the results of the experiments are shown in Table 3. We obtained an mIoU metric of 0.782 on the original CNN-based network, which then increased to 0.792 after embedding the RB module in it. We obtained an mIoU of 0.821 on Kvasir-SEG after using the vision transformer as the backbone, which proves that transformer added as a CNN has a more significant effect than the original CNN. The final mIoU metric of our method on the Kvasir-SEG dataset is 0.893.

Method	mIoU
Encoder + Decoder	0.782
Encoder + RB + Decoder	0.792
Trans + Decoder	0.821
Trans + RB + Decoder	0.834
Trans + GCE + Decoder	**0.842**
Trans + GCE + RB + Decoder	**0.867**
Trans + GCE + MHSA + RB + Decoder (CTransNet)	**0.893**

"Trans" represents vision transformer.

Table 3. mIoU with different setting on Kvasir-SEG dataset

6 CONCLUSION

In this paper, the proposed CTransNet effectively combines CNN with the self-attention mechanism in Transformer to improve the performance of medical image segmentation. This hybrid framework does not require Transformer to be pre-trained on large-scale datasets, where self-attention can effectively capture different levels of long-range information. We believe that this design will help design richer Transformer models that are more suitable for medical image segmentation tasks; in addition, the excellent ability to handle long-range sequences in CTransNet opens up the possibility of migration to other downstream tasks. In the future, we will be working on the task of analysing medical image segmentation from a semi-supervised or weakly supervised perspective. This will give us access to fewer datasets and a more scientific approach to deep learning, and we will also be working on the segmentation of small medical targets.

Acknowledgements

The authors express their gratitude to the reviewers for their valuable feedback and recommendations, which have significantly contributed to improving the quality of the manuscript. The authors would also like to acknowledge the support and assistance provided by their colleagues and students in the laboratory during the course of this research.

REFERENCES

[1] Beltagy, I.—Peters, M. E.—Cohan, A.: Longformer: The Long-Document Transformer. 2020, arXiv: 2004.05150.

[2] Xie, Y.—Zhang, J.—Shen, C.—Xia, Y.: CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 171–180, doi: 10.1007/978-3-030-87199-4_16.

[3] Wolf, T.—Debut, L.—Sanh, V.—Chaumond, J.—Delangue, C.—Moi, A.—Cistac, P.—Rault, T.—Louf, R.—Funtowicz, M. et al.: Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.

[4] Yu, S.—Ma, K.—Bi, Q.—Bian, C.—Ning, M.—He, N.—Li, Y.—Liu, H.—Zheng, Y.: Mil-Vt: Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 45–54, doi: 10.1007/978-3-030-87237-3_5.

[5] Zhang, Z.—Sun, B.—Zhang, W.: Pyramid Medical Transformer for Medical Image Segmentation. 2021, doi: 10.48550/arXiv.2104.14702.

[6] Dai, Y.—Gao, Y.—Liu, F.: Transmed: Transformers Advance Multi-Modal Medical Image Classification. Diagnostics, Vol. 11, 2021, No. 8, Art. No. 1384.

[7] Shaw, P.—Uszkoreit, J.—Vaswani, A.: Self-Attention with Relative Position Representations. 2018, arXiv: 1803.02155.

[8] Tsai, A.—Yezzi, A.—Wells, W.—Tempany, C.—Tucker, D.—Fan, A.—Grimson, W. E.—Willsky, A.: A Shape-Based Approach to the Segmentation of Medical Imagery Using Level Sets. IEEE Transactions on Medical Imaging, Vol. 22, 2003, No. 2, pp. 137–154, doi: 10.1109/TMI.2002.808355.

[9] Held, K.—Kops, E. R.—Krause, B. J.—Wells, W. M.—Kikinis, R.—Muller-Gartner, H. W.: Markov Random Field Segmentation of Brain MR Images. IEEE Transactions on Medical Imaging, Vol. 16, 1997, No. 6, pp. 878–886, doi: 0.1016/j.eswa.2019.05.038.

[10] Li, X.—Chen, H.—Qi, X.—Dou, Q.—Fu, C. W.—Heng, P. A.: H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. IEEE Transactions on Medical Imaging, Vol. 37, 2018, No. 12, pp. 2663–2674, doi: 10.1109/TMI.2018.2845918.

[11] Ronneberger, O.—Fischer, P.—Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[12] Xiao, X.—Lian, S.—Luo, Z.—Li, S.: Weighted Res-Unet for High-Quality Retina Vessel Segmentation. 2018 9th International Conference on Information Technology in Medicine and Education (ITME), IEEE, 2018, pp. 327–331, doi: 10.1109/ITME.2018.00080.

[13] Zhou, Z.—Rahman Siddiquee, M. M.—Tajbakhsh, N.—Liang, J.: Unet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.

[14] Huang, H.—Lin, L.—Tong, R.—Hu, H.—Zhang, Q.—Iwamoto, Y.—Han, X.—Chen, Y. W.—Wu, J.: Unet 3+: A Full-Scale Connected Unet for Medical Image Segmentation. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 1055–1059.

[15] Çiçek, Ö.—Abdulkadir, A.—Lienkamp, S. S.—Brox, T.—Ronneberger, O.: 3d U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 424–432, doi: 10.1007/978-3-319-46723-8_49.

[16] Milletari, F.—Navab, N.—Ahmadi, S. A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571, doi: 10.1109/3DV.2016.79.

[17] Vaswani, A.—Shazeer, N.—Parmar, N.—Uszkoreit, J.—Jones, L.—Gomez, A. N.—Kaiser, Ł.—Polosukhin, I.: Attention Is All You Need. Advances in Neural Information Processing Systems, Vol. 30, 2017.

[18] Devlin, J.—Chang, M. W.—Lee, K.—Toutanova, K.: Bert: Pre-Training

of Deep Bidirectional Transformers for Language Understanding. 2018, arXiv: 1810.04805.

[19] Dosovitskiy, A.—Beyer, L.—Kolesnikov, A.—Weissenborn, D.—Zhai, X.—Unterthiner, T.—Dehghani, M.—Minderer, M.—Heigold, G.—Gelly, S. et al.: An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. 2020, arXiv: 2010.11929.

[20] Touvron, H.—Cord, M.—Douze, M.—Massa, F.—Sablayrolles, A.—Jégou, H.: Training Data-Efficient Image Transformers & Distillation Through Attention. International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[21] Wang, W.—Xie, E.—Li, X.—Fan, D. P.—Song, K.—Liang, D.—Lu, T.—Luo, P.—Shao, L.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578, doi: 10.1109/ICCV48922.2021.00061.

[22] Han, K.—Xiao, A.—Wu, E.—Guo, J.—Xu, C.—Wang, Y.: Transformer in Transformer. Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 15908–15919.

[23] Liu, Z.—Lin, Y.—Cao, Y.—Hu, H.—Wei, Y.—Zhang, Z.—Lin, S.—Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[24] Wang, X.—Girshick, R.—Gupta, A.—He, K.: Non-Local Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[25] Chen, J.—Lu, Y.—Yu, Q.—Luo, X.—Adeli, E.—Wang, Y.—Lu, L.—Yuille, A. L.—Zhou, Y.: Transunet: Transformers Make Strong Encoders for Medical Image Segmentation. 2021, arXiv: 2102.04306.

[26] Li, Z.—Chen, G.—Zhang, T.: A CNN-Transformer Hybrid Approach for Crop Classification Using Multitemporal Multisensor Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 13, 2020, pp. 847–858, doi: 10.1109/JSTARS.2020.2971763.

[27] Luo, X.—Hu, M.—Song, T.—Wang, G.—Zhang, S.: Semi-Supervised Medical Image Segmentation via Cross Teaching Between CNN and Transformer. 2021, arXiv: 2112.04894.

[28] Weng, W.—Zhang, Y.—Xiong, Z.: Event-Based Video Reconstruction Using Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2563–2572, doi: 10.1109/ICCV48922.2021.00256.

[29] Liang, J.—Cao, J.—Sun, G.—Zhang, K.—Van Gool, L.—Timofte, R.: Swinir: Image Restoration Using Swin Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1833–1844, doi: 10.1109/ICCVW54120.2021.00210.

[30] Schlemper, J.—Oktay, O.—Schaap, M.—Heinrich, M.—Kainz, B.—Glocker, B.—Rueckert, D.: Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. Medical Image Analysis, Vol. 53, 2019,

pp. 197–207, doi: 10.1016/j.media.2019.01.012.

[31] Valanarasu, J. M. J.—Oza, P.—Hacihaliloglu, I.—Patel, V. M.: Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 36–46, doi: 10.1007/978-3-030-87193-2_4.

[32] Hatamizadeh, A.—Tang, Y.—Nath, V.—Yang, D.—Myronenko, A.—Landman, B.—Roth, H. R.—Xu, D.: Unetr: Transformers for 3d Medical Image Segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584, doi: 10.1109/WACV51458.2022.00181.

[33] Liu, Y.—Hu, J.—Kang, X.—Luo, J.—Fan, S.: Interactformer: Interactive Transformer and CNN for Hyperspectral Image Super-Resolution. IEEE Transactions on Geoscience and Remote Sensing, Vol. 60, 2022, pp. 1–15, doi: 10.1109/TGRS.2022.3183468.

[34] Zhang, Y.—Liu, H.—Hu, Q.: Transfuse: Fusing Transformers and Cnns for Medical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 14–24, doi: 10.1007/978-3-030-87193-2_2.

[35] Wang, W.—Chen, C.—Ding, M.—Yu, H.—Zha, S.—Li, J.: Transbts: Multimodal Brain Tumor Segmentation Using Transformer. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 109–119, doi: 10.1007/978-3-030-87193-2_11.

[36] Alom, M. Z.—Hasan, M.—Yakopcic, C.—Taha, T. M.—Asari, V. K.: Recurrent Residual Convolutional Neural Network Based on U-Net (r2u-Net) for Medical Image Segmentation. 2018, arXiv: 1802.06955.

[37] Shin, S. Y.—Lee, S.—Yun, I. D.—Lee, K. M.: Deep Vessel Segmentation by Learning Graphical Connectivity. Medical Image Analysis, Vol. 58, 2019, Art. No. 101556, doi: 10.1016/j.media.2019.101556.

[38] Jaeger, P. F.—Kohl, S. A.—Bickelhaupt, S.—Isensee, F.—Kuder, T. A.—Schlemmer, H.—Maier-Hein, K. H.: Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection. ML4H Workshop, PMLR, 2020, pp. 171–183.

[39] Oktay, O.—Schlemper, J.—Folgoc, L. L.—Lee, M.—Heinrich, M.—Misawa, K. et al.: Attention U-Net: Learning Where to Look for the Pancreas. 2018, arXiv:1804.03999.

[40] Azad, R.—Asadi-Aghbolaghi, M.—Fathy, M.—Escalera, S.: Bi-Directional ConvLSTM U-Net with Densley Connected Convolutions. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 406–415, doi: 10.1109/ICCVW.2019.00052.

[41] Gu, Z.—Cheng, J.—Fu, H.—Zhou, K.—Hao, H.—Zhao, Y. et al.: CE-Net: Context Encoder Network for 2D Medical Image Segmentation. IEEE Trans. Med. Imaging, Vol. 38, 2019, No. 10, pp. 2281–2292, doi: 10.1109/TMI.2019.2903562.

[42] Ding, J.—Zhang, Z.—Tang, J.—Guo, F.: A Multichannel Deep Neural Network for Retina Vessel Segmentation via a Fusion Mechanism. Frontiers in Bioengineering and Biotechnology, 2021, 663 pp., doi: 10.3389/fbioe.2021.697915.

[43] Zhang, Z.—Liu, Q.—Wang, Y.: Road Extraction by Deep Residual U-Net. IEEE Geoscience and Remote Sensing Letters, Vol. 15, 2018, No. 5, pp. 749–753, doi: 10.1109/LGRS.2018.2802944.

[44] Jha, D.—Smedsrud, P. H.—Johansen, D.—de Lange, T.—Johansen, H. D.—Halvorsen, P.—Riegler, M. A.: A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, Conditional Random Field and Test-Time Augmentation. IEEE Journal of Biomedical and Health Informatics, Vol. 25, 2021, No. 6, pp. 2029–2040.

[45] Jha, D.—Smedsrud, P. H.—Riegler, M. A.—Johansen, D.—De Lange, T.—Halvorsen, P.—Johansen, H. D.: Resunet++: An Advanced Architecture for Medical Image Segmentation. 2019 IEEE International Symposium on Multimedia (ISM), IEEE, 2019, pp. 225–2255, doi: 10.1109/ISM46123.2019.00049.

[46] Chen, L. C.—Zhu, Y.—Papandreou, G.—Schroff, F.—Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.

[47] Tomar, N. K.—Jha, D.—Ali, S.—Johansen, H. D.—Johansen, D.—Riegler, M. A.—Halvorsen, P.: DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation. International Conference on Pattern Recognition, Springer, 2021, pp. 307–314, doi: 10.1007/978-3-030-68793-9_23.

[48] Srivastava, A.—Jha, D.—Chanda, S.—Pal, U.—Johansen, H. D.—Johansen, D.—Riegler, M. A.—Ali, S.—Halvorsen, P.: Msrf-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation. IEEE Journal of Biomedical and Health Informatics, Vol. 26, 2021, No. 5, pp. 2252–2263, doi: 10.1109/JBHI.2021.3138024.

Zhixin Zhang graduated from the Tianjin University of Technology. He is currently Lecturer in the College of Information Engineering, Tianjin University of Commerce. His research interests include image recognition and intelligence computing.

Shuhao Jiang received his Master of Engineering in the Tianjin Normal University, Ph.D. in Tianjin University. He is currently Professor in the College of Information Engineering, Tianjin University of Commerce. His research interests include intelligence computing and Natural Language Processing.

Xuhua Pan graduated from the Jilin University. He is currently Professor in the College of Information Engineering, Tianjin University of Commerce. His research interests include intelligence computing and data handling.