

# Transformer and CNN combined generative adversarial network for segmentation of placental tissue in MR images

YE Zhengjie<sup>1</sup>, WANG Yutao<sup>2</sup>, XU Jian<sup>3</sup>, JIN Wei<sup>1\*</sup>

( 1.Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China;

2.Affiliated Hospital of Medical School, Ningbo University, Ningbo 315020, China;

3.Ningbo Women & Children's Hospital, Ningbo 315012, China )

**Abstract:** Accurate segmentation of placental tissue in magnetic resonance (MR) images is of great significance for the study of pregnancy and childbirth complications. However, manual annotation by radiologists is difficult to ensure the accuracy and objectivity of the results, and it is time-consuming and labor-intensive. To develop a deep learning model for automated placental tissue segmentation in MR images, we propose a generative adversarial network (TCGANet) that combines transformer and convolution neural network (CNN) to realize the segmentation purposes. It combines the feature embedding module with the skip connection which avoids the information loss caused by the traditional feature fusion method. On the basis of this technique, the context extractor module is introduced, and the self-attention mechanism of the transformer is adopted to capture the global dependencies, which can effectively represent the global and local information of MR images. In addition, to improve the precision of placental tissue edge segmentation, the generative adversarial architecture is introduced to supervise the generative network of placental tissue segmentation by the discriminator network. The results show that the proposed model is superior to other comparison methods in terms of quantitative metrics and boundary positioning accuracy, the accuracy of which reads  $0.993\pm0.003$ , sensitivity reaches  $0.903\pm0.093$ , specificity is  $0.996\pm0.003$ , and Dice similarity coefficient is found to be  $0.861\pm0.141$ . Ablation experiments on different structural units of the TCGANet have validated the network structure design, and most of the performance metrics are far better than those of the existing methods ( $P<0.05$ ). The proposed TCGANet enables automated and accurate segmentation of placental tissue in MR images.

**Key words:** deep learning; placental tissue segmentation; adversarial network; MRI; attention mechanism

**CLC number:** TP391.4

**Document code:** A

**Article ID:** 1001-5132 ( 2023 ) 01-0022-13

## 1 Introduction

The placenta is the critical organ for material exchange between the fetus and the mother, and the fetus relies on it to absorb nutrients from the mother. Placental abnormalities<sup>[1]</sup> are common obstetric diseases

and may cause serious obstetric complications. Magnetic resonance (MR) imaging has the advantages of a large field of view and good soft tissue contrast and is an important complement to ultrasound<sup>[2]</sup>. T2 weighted image (T2WI) clearly distinguish the placental tissue from the others, so the research in this paper is

收稿日期: 2022-07-19.

宁波大学学报(理工版)网址: <http://journallg.nbu.edu.cn/>

基金项目: 浙江省自然科学基金(LY20H180003); 宁波市自然科学基金(2019A610104); 宁波市公益类科技计划项目(202002N3104).

第一作者: 叶正洁(1996-), 女, 浙江丽水人, 在读硕士研究生, 主要研究方向: 医学影像. E-mail: 1303698906@qq.com

\*通信作者: 金炜(1969-), 男, 浙江金华人, 教授, 主要研究方向: 数字图像处理. E-mail: xyjw1969@126.com

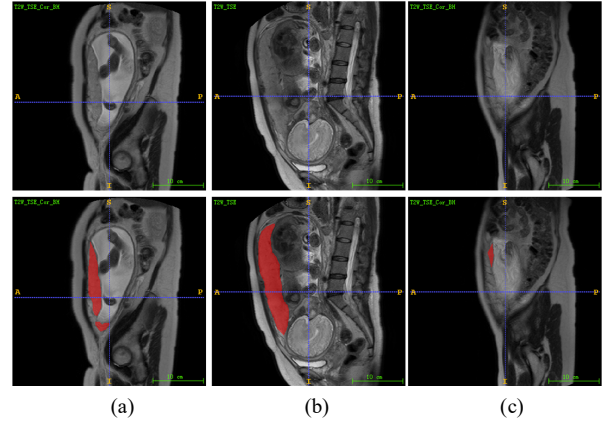
based on T2WI<sup>[3]</sup>. Accurate segmentation of placental tissue is essential for diagnosing placental abnormalities<sup>[4]</sup>. Traditionally, placental tissue segmentation mainly relies on manual labeling by radiologists, which is difficult to ensure the accuracy and objectivity of labeling results, and it is time-consuming and labor-intensive. Therefore, it is urgent to design an effective and reliable automatic segmentation method of placental tissue in MR images.

In recent years, convolution neural networks (CNNs) and other deep learning-based algorithms have developed rapidly, and have been widely used in the fields of medical image segmentation<sup>[5]</sup>, classification<sup>[6-7]</sup>, and detection<sup>[8]</sup>. U-Net<sup>[9]</sup> is one of the most popular networks in the field of medical image segmentation, and subsequent improvements based on U-Net<sup>[10-11]</sup> are also widely used. Some studies have achieved great performance in medical image segmentation such as that of liver<sup>[12]</sup>, brain<sup>[13]</sup>, prostate<sup>[14]</sup>, and other organs.

Although CNNs have been widely used in the field of medical image segmentation because of their excellent feature representation capabilities, they lack the ability to model long-range dependencies present in an image. More precisely, each convolution kernel only has a local subset of pixels in the whole image. Therefore, to obtain global information, these models stack multiple convolution layers and expand the receptive field through downsampling, which has limitations<sup>[15]</sup>. Some works have been done to model the long-range dependencies of CNNs using image pyramids<sup>[16]</sup> and atrous convolution<sup>[17]</sup>. Although these deep learning models could gain accurate segmentation masks, there are also varieties of challenges in placental tissue segmentation. Three major challenges to placental tissue segmentation in MR imaging are illustrated in Fig. 1.

For placental tissue regions with discontinuous or large spans, as shown in Fig. 1(a) and (b), learning long-range dependencies between the pixels corresponding to the mask is helpful to improve the performance of placental segmentation. For the small placental tissue area, as shown in Fig. 1(c), more attention should be paid to local information. However, the placental tissue

structures are irregular, complex, and multi-scale. Perceiving these placental tissue features with fixed regions could be inadequate.



**Fig. 1 Three challenging cases, including (a) discontinuity of placental tissue area, (b) large span of placental tissue area, (c) small placental tissue area**

Therefore, local and global information should be comprehensively considered in placental tissue segmentation, while most medical segmentation networks based on CNNs ignore the long-range contextual information. Transformer have recently revolutionized most of natural language processing (NLP) tasks like machine translation, question answering and document classification. The main reason for the success of transformers is their ability to learn long-range dependencies between the input tokens. Therefore, we believe that the combination of CNN and transformer is expected to achieve more accurate segmentation.

To address these challenges, this paper proposes TCGANet, which is based on a generative adversarial network. In this work, we aim to alleviate the aforementioned challenges from different perspectives. On the one hand, to better segment the placental tissues with different sizes and shapes, we made full use of global and local information. On the other hand, to accurately define the boundary of placental tissue, we used an adversarial network to monitor the generation network of placental tissue segmentation. The main contributions of this work can be summarized as follow:

- (1) A deep network for automatic segmentation of

placental tissue in MR images was constructed. Through the combination of the feature embedding module and skipping connection, the network embedded the high-level features into the low-level features, alleviating the information loss caused by the traditional feature fusion method based on skip connection, and improving the accuracy of placental tissue segmentation.

(2) In the network architecture, a context feature extractor was introduced. This module can make full use of the ability of CNN to obtain local features and advantages of the transformer to learn long-term dependence, thus facilitating accurate segmentation of placental tissues with regional discontinuity and large size differences.

(3) Transformer and convolution neural network (CNN) were combined to form a generative adversarial network (TCGANet). Based on this architecture, TCGANet uses an adversarial network to supervise the generation network of placental tissue segmentation, thus improving the accuracy of placental edge localization.

The paper is organized as follows: Section 2 provides the related work. Section 3 describes the TCGANet network architecture proposed in this paper in detail. The experiment and results are introduced in section 4. Finally, the conclusive remarks and future work are given in sections 5–6.

## 2 Related Work

### 2.1 CNNs-based approaches

An end-to-end network based on CNN is widely used in image segmentation. Following the initial breakthrough made by FCN<sup>[18]</sup> in image segmentation, Li et al<sup>[19]</sup> designed a 3D multi-scale network for the segmentation of multimodal intervertebral disc MR images based on FCN. U-Net<sup>[9]</sup> is generally regarded as the baseline for medical image segmentation tasks, and various improvements have also been made for different segmentation tasks<sup>[20]</sup>. Shahedi et al<sup>[21]</sup> proposed an improved U-Net for uterine and placental tissue segmentation in MR images. Despite the significant

progress made by these methods, they are not sufficient to address the challenges of placental tissue segmentation.

### 2.2 GAN-based approaches

Generative adversarial networks (GAN) includes generator network and discriminator network<sup>[22]</sup>. Luc et al<sup>[23]</sup> applied GAN to the field of semantic segmentation. GAN and its extensions solved many well-known challenging medical image problems<sup>[24]</sup>. In the field of medical image segmentation, inspired by the classic GAN, SegAN was proposed by Xue et al<sup>[25]</sup> as an end-to-end neural network for skin damage segmentation. Dai et al<sup>[26]</sup> proposed SCAN to segment lungs and hearts in chest X-ray images. Accurate segmentation results can also be obtained by using only limited training data. GAN-based methods are constantly being explored and improved, which provide the possibility to further improve the accuracy of segmentation.

### 2.3 Transformer-based approaches

Transformers<sup>[27]</sup> have become the model of choice in natural language processing (NLP). Enlightened by NLP successes, scholars began to explore the application of transformers in tasks in the field of computer vision, such as object detection in DETR<sup>[28]</sup>. Vision transformer (ViT)<sup>[29]</sup> achieved excellent performance in image classification, and more than this, it laid the foundation for semantic segmentation. Valanarasu et al<sup>[15]</sup> designed a medical transformer (MedT) for medical image segmentation, which significantly improved the segmentation performance. In addition, Chen et al<sup>[30]</sup> used CNNs and transformer to build a more powerful hybrid encoder, confirming that the transformer's strong long-range sequence modeling ability can make up for some of the deficiencies of CNNs. Many scholars have been combining transformer structure with convolution operation to achieve better performance<sup>[31-32]</sup> and exploring the performance of transformers in segmentation is one of current research trends. Several recent studies have introduced hybrid transformer-CNN models for MRI processing tasks such as reconstruction and synthesis that are highly related to

the proposed architecture<sup>[33-34]</sup>.

Therefore, this paper draws on the above-mentioned existing state-of-the-art methods and proposes TCGANet for automatic placental tissue segmentation in MR images.

### 3 Materials and Methods

In this section, the proposed method and the related experimental setup and data preparation are described in detail here.

#### 3.1 Dataset description and preprocessing

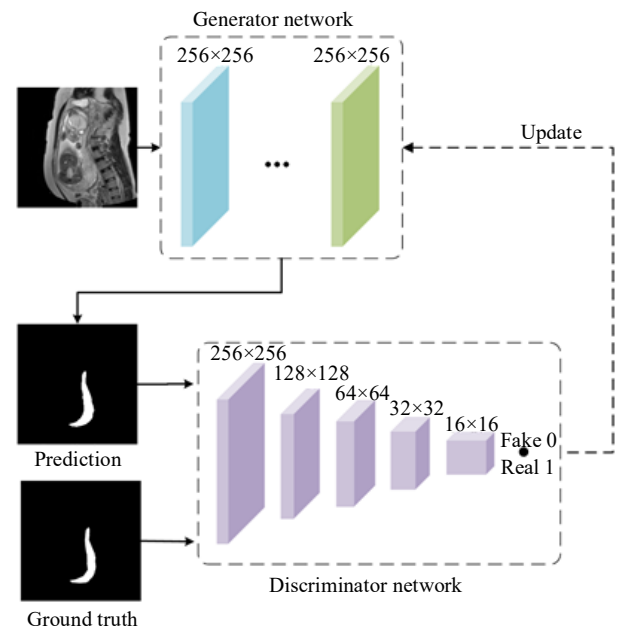
With the approval of the ethics committee, the data provided by the Affiliated Hospital of Medical School of Ningbo University and Ningbo Women and Children's Hospital were used in this paper. The imaging equipment of the Affiliated Hospital of Medical School of Ningbo University is Ge signa twin speed 1.5T superconducting dual gradient magnetic resonance scanner with an 8-channel body phased array coil. The imaging equipment of Ningbo Women and Children's Hospital is Philips Achieva Noval Dual 1.5T superconducting dual gradient magnetic resonance scanner, using a 16-channel body phased array coil. The scanned slice thickness is 4.5 mm, the slice spacing is 2 mm, and the matrix is  $256 \times 256$ . A total of 40 placental MR images cases were collected, of which 27 were from the Ningbo Women and Children's Hospital and 13 were from the Affiliated Hospital of Medical School of Ningbo University. The placental tissue of each case is manually labeled by an experienced radiologist. According to the image quality, a total of 572 sagittal slices with T2-sequential containing placental tissues were selected. K-fold cross-validation can prevent overfitting problems caused by overly complex models and is a method that can improve the generalization ability of models. In order to maximize the training process of the model, improve the reliability of the test results and avoid over-fitting and under-fitting, this experiment uses the 5-fold cross-validation method. To avoid data leakage, all slices are divided into training

and test sets by case.

Due to the scarcity of available data, data augmentation is critical for the analysis of medical images, and augmenting the dataset can also reduce the risk of overfitting. Standard data augmentation techniques are utilized, including image translation, rotation and scaling, vertical flipping, and horizontal flipping.

#### 3.2 TCGANet architecture

The overall structure is shown in Fig. 2, which is composed of the generator network (G) and discriminator network (D).



**Fig. 2 Architecture of proposed TCGANet, which includes generator network and discriminator network**

During the training stage, G and D are trained alternately. When training the D, only the parameters of the D are updated, and the same scheme is applied to the G. After training, the segmentation results can be obtained from the model.

##### 3.2.1 Generator network

The backbone of the generator network is constructed based on the basic U-Net, as shown in Fig. 3. To mitigate the information loss caused by traditional skip connections, we are combining the feature embedding module and skip connections. The context extractor module is designed to improve the limitation

that convolution can only obtain local features by capturing features of different scales and long-term context information. The details are as follows:

#### (1) Feature encoder module

In this paper, resnet<sup>[35]</sup> is chosen as the encoder. It was found that the pre-trained resnet<sup>[34]</sup> can achieve better performance through experiments. The first four feature extraction modules are retained, and each module is shown in Fig. 3(a). The input of the first layer can be defined as  $H \times W \times C$ ,  $H \times W$  is the input image size,  $C$  is the number of channels. The output is a three-dimensional tensor  $h \times w \times c$ ,  $h \times w$  is the spatial dimension of the feature map,  $c$  is the number of channels of the feature map. Networks such as U-Net usually directly use skip connection to combine the low-level features obtained by the encoder with the upsampled high-level features. The fusion is as follows:

$$y_i = [U(y_{i+1}), x_i],$$

where  $x_i$  is the output feature of the  $i$ -th layer of the encoder.  $[\cdot]$  denotes concatenation operation. This fusion method has drawbacks: First of all, due to significant differences in semantic information and resolution contained in low-level and high-level features,

it is difficult to be fused directly. Secondly, a shallower network pays more attention to detailed information, including encoded clearer semantic boundaries. Instead, a deeper network pays more attention to semantic information and can describe the object more accurately. In short, a direct skip connection will cause some useful information loss. To solve the above problems, inspired by [36], this paper uses the feature embedding module and uses  $F(\cdot)$  to express this operation. This feature fusion can be expressed as:

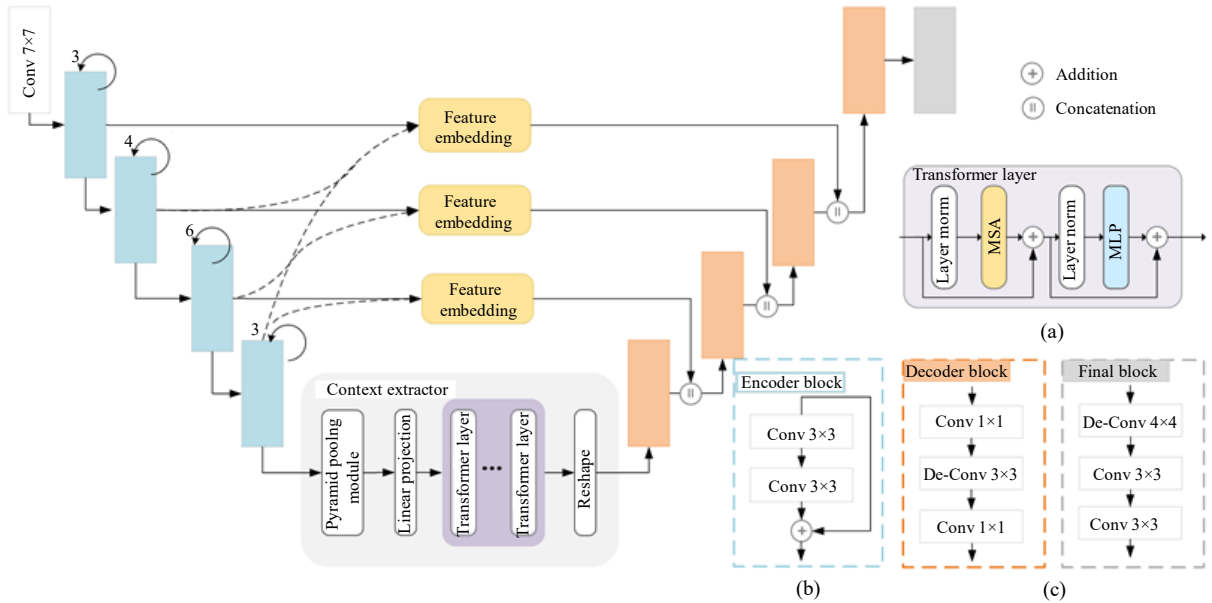
$$y_i = [U(y_{i+1}), F(x_i, x_{i+1}, \dots, x_n)],$$

where  $U(\cdot)$  denotes the upsampling operation and  $n$  denotes the number of feature layers. Its detailed description is shown in Fig. 4, where the “ $\times$ ” symbol represents the multiplication operation of elements. Embedding high-level features into low-level features are beneficial to supplementing low-level features.

As shown in Fig. 3, the feature embedding module is applied to the first three layers of the encoder.

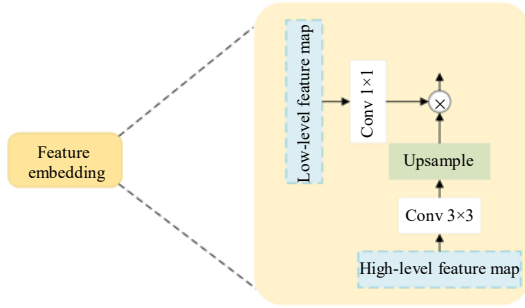
#### (2) Context extractor module

Context extractor module is designed, as shown in Fig. 3, which is mainly composed of pyramid pooling module<sup>[16]</sup> and  $L$  layers transformer. For placental



**Fig. 3 Architecture of proposed generator network, which consists of three parts: feature encoder module, context extractor module, and feature decoder module. Several key network blocks from different parts and some main notations are illustrated on the right**

tissues with large size differences, we used a pyramid pooling module to capture and aggregate features of different scales. For placental tissues with a large span, the transformer is introduced to capture long-term dependence using a self-attention mechanism. This module fully establishes the correlation of each pixel and integrates local features with global context information.



**Fig. 4 The structure of the feature embedding module**

Context information is extracted from different perspectives and scales through the pyramid pooling module. Features of four different scales are merged, using the average pooling of sizes  $1 \times 1$ ,  $2 \times 2$ ,  $3 \times 3$  and  $6 \times 6$ . In this way, more spatial information can be obtained, which contributes to the segmentation of placental tissue with large differences in size. The output feature map can be expressed as  $f \in R^{h \times w \times d}$ .

Because the global information is beneficial to the segmentation of the placental tissue with a large area, the transformer is introduced to further capture the global information using self-attention mechanisms. The standard transformer expects a sequence as input. Thus, the feature map  $F \in R^{h \times w \times d}$  is divided into patches of size  $P \times P$  and flattened to obtain  $N \times P^2$  matrix, where  $N = hw / P^2$  is the input sequence length. Further mapping vectorized patches into a latent  $d$ -dimensional embedding space using a linear projection function. To encode the location information which is vital in placental tissue segmentation task, we add a 1D learnable position embedding  $E_{\text{pos}} \in R^{N \times d}$  to the patch embedding to retain positional information as follows:

$$z_0 = F_{\text{LP}}(f) + E_{\text{pos}},$$

where  $z_0 \in R^{N \times d}$  refers to the feature embeddings, and

$F_{\text{LP}}(\cdot)$  is the linear projection operation. The transformer encoder is composed of  $L$  transformer layers, each of them has a standard architecture as shown in Fig. 3(b), which consists of a multihead self-attention (MSA) and multi-layer perceptron (MLP) blocks. Therefore the output of the  $\ell$ -th ( $\ell \in [1, 2, \dots, L]$ ) layer can be written as follows:

$$z'_\ell = F_{\text{MSA}}(F_{\text{LN}}(z_\ell)) + z_{\ell-1},$$

$$z_\ell = F_{\text{MLP}}(F_{\text{LN}}(z'_\ell)) + z'_\ell,$$

where  $F_{\text{LN}}(\cdot)$  denotes the layer normalization operator and  $z_\ell$  is the output of  $\ell$ -th transformer layer. MSA is composed of  $m$  parallel self-attention (SA) heads, and the output is defined as:

$$F_{\text{MSA}}(z) = [F_{\text{SA}_1}(z), F_{\text{SA}_2}(z), \dots, F_{\text{SA}_m}(z)]W^O,$$

where  $W^O$  represents the learnable weight matrices of different SA heads. SA is a parameterized function that learns the similarity between two elements in the input sequence ( $z$ ) and their set of the query ( $q$ ) and key ( $k$ ) representations. SA is formulated as:

$$F_{\text{SA}}(z) = F_{\text{Softmax}}(qk^T / \sqrt{d_h})v,$$

where  $v$  denotes the values in the input sequence and  $d_h = d / m$  is a scaling factor. Finally, to fit the input dimension of the decoder, the sequence data  $z_\ell \in R^{N \times d}$  is projected back to a standard 3D feature map  $F \in R^{h \times w \times d}$  through the reshape operation.

### (3) Feature decoder module

The feature decoder module is made up of four decoder blocks as shown in Fig. 3(c), designed to recover high-resolution information and the feature map size from more abstract semantic features. The bottleneck structure is used to construct the decoder block. Each decoder block contains two pointwise convolutions to reduce the number of channels. Between the two pointwise convolutions, there is a layer of  $3 \times 3$  deconvolution, which could learn a self-adaptive mapping to restore features with more detailed information. Ultimately, the feature decoder module outputs a mask of the same size as the original image.

#### 3.2.2 Discriminator network

The edge definition of placental tissue is essential

for placental tissue segmentation. The discriminative network can highlight the difference of edge distribution between segmented and real maps, enhance the perception of placental tissue edges, guide the segmentation network training, and further improve the accuracy of placental tissue boundary definition. The discriminator takes two input combinations, one set of original image and labeled image, and one set of original image and generator placenta segmentation image. The discriminator network distinguishes between ground truth and predicted mask, and provides a trainable loss function for the generator network, to supervise its training. Each block of the discriminator network includes 2 layers of convolution with a kernel size of  $3 \times 3$ , a stride of 1, and a padding of 1. The last layer of the discriminator network is activated using the sigmoid function.

### 3.3 Objective function

Assume that  $x$  is an input image and  $y$  is the ground truth.  $G(x)$  is the segmentation result of generator network,  $D(x, y)$  are the probability that the input is ground truth, and  $D(x, G(x))$  indicates the probability that the input is predicted mask. By using generator network to minimize the objective function, and discriminator network to maximize the objective function, the following function is optimized:

$$G^* = \arg \min_G \left[ \max_D \left[ E_{x, y \sim P_{\text{data}}(x, y)} \log D(x, y) + E_{x \sim P_{\text{data}}(x)} \log(1 - D(x, G(x))) \right] \right].$$

To better supervise the generator network and further improve the accuracy of placental tissue segmentation, a binary cross-entropy loss function is added to the generator network:

$$G_{\text{seg}} = E_{x, y \sim p_{\text{data}}} (-y \log(G(x)) - (1-y) \log(1-G(x))).$$

The final objective function is defined as:

$$G_{\text{loss}} = G^* + G_{\text{seg}}.$$

## 4 Experiments and Results

### 4.1 Implementation details

The proposed method is based on the Windows 10

system and implemented by using Python 3.6 and Pytorch 1.7.1. The graphics card used is NVIDIA GeForce GTX 1080Ti GPU. Our experiments use a batch size of 8, use Adam optimization, the learning rate is  $10^{-4}$ , the number of layers of the transformer is set to 12, the embedding dim is 1 024, and the model is trained for 100 epochs. For all the comparison methods in this study, training and testing are carried out according to their papers and codes, and the other settings are the same as in this paper.

### 4.2 Evaluation metrics

In this paper, general indicators in the field of image segmentation are applied to measure the performance of different segmentation methods<sup>[37]</sup>. In this study, True Positive ( $P_T$ ) is the correctly segmented placental tissue pixels, and False Positive ( $P_F$ ) is the wrongly segmented placental tissue pixels, also called over-segmented pixels. Similarly, True Negative ( $N_T$ ) means the correctly segmented non-placental tissue pixels, and False Negative ( $N_F$ ) is the pixels of placental tissue that are not segmented. Based on these definitions, the following evaluation indicators are introduced respectively. The mathematical expressions of the six metrics:  $M_{\text{ACC}}$ ,  $M_{\text{SEN}}$ ,  $M_{\text{SPE}}$ ,  $M_{\text{PPV}}$ ,  $M_{\text{DSC}}$ , and  $M_{\text{JSC}}$  are presented below:

$$M_{\text{ACC}} = (P_T + N_T) / (P_T + N_T + P_F + N_F),$$

$$M_{\text{SEN}} = P_T / (P_T + N_F),$$

$$M_{\text{SPE}} = N_T / (N_T + P_F),$$

$$M_{\text{PPV}} = P_T / (P_F + P_T),$$

$$M_{\text{DSC}} = 2P_T / (P_F + 2P_T + N_F),$$

$$M_{\text{JSC}} = P_T / (P_T + N_F + P_F).$$

Accuracy ( $M_{\text{ACC}}$ ) is the entire accuracy of segmentation of placental tissue pixels and non-placental tissue pixels. Sensitivity ( $M_{\text{SEN}}$ ) is applied to evaluate the proportion of correctly segmented placental tissue pixels in the ground truth placental tissue pixels. Specificity ( $M_{\text{SPE}}$ ) aims to measure the ratio of how many non-placental tissue pixels from the non-placental tissue pixels ground truth are classified to be true. Positive predictive value ( $M_{\text{PPV}}$ ) indicates the proportion of placental tissue pixels in the prediction result that

belong to the ground truth. Dice similarity coefficient ( $M_{DSC}$ ) is a commonly used evaluation index in medical image segmentation. Jaccard similarity coefficient ( $M_{JSC}$ ) is defined as the intersection of the segmentation result and ground truth divided by their union.

### 4.3 Results

In this section, the results of the ablation experiment are presented, and the qualitative and quantitative analysis results of the model proposed in this paper and the state-of-the-art methods are shown in detail.

#### 4.3.1 Ablation study

To verify the effectiveness of each part of the proposed network in the task of segmenting the placental tissue, and to further discuss the contribution of each part, the following ablation experiments in this paper are performed.

We conducted the ablation study on placental tissue segmentation by evaluating three key components of TCGANet (discriminator, feature embedding module, and context extractor module). The detailed experimental results are presented in Table 1.

**Table 1 Ablation study on different components of TCGANet**

| Method              | $M_{DSC}/\%$ |
|---------------------|--------------|
| TCGANet w/o D, F, C | 71.48        |
| TCGANet w/o D, F    | 72.68        |
| TCGANet w/o D, C    | 77.03        |
| TCGANet w/o F, C    | 82.56        |
| TCGANet w/o C       | 84.68        |
| TCGANet w/o F       | 85.03        |
| TCGANet             | 86.19        |

Notes: D is discriminator; F is feature embedding module; C is context extractor module; w/o is without. The same below.

For ease of presentation, we signify discriminator, feature embedding module, and context extractor module as D, F, and C, respectively. As shown in Table 1, when TCGANet without (w/o) D, the  $M_{DSC}$  is less than 80%, which shows that D effectively improves the overall segmentation performance. Similarly, F and C also contribute to placental tissue segmentation, because the  $M_{DSC}$  of TCGANet w/o F, C is 82.56%. Compared

with TCGANet, the  $M_{DSC}$  of TCGANet w/o C dropped from 86.19% to 84.68%. C effectively utilizes global and local information, which shows that the designed C can help achieve better segmentation. The  $M_{DSC}$  of TCGANet w/o F is 85.03%, which proves that a suitable feature fusion strategy can further improve the segmentation performance of the placental tissue. It also means that correctly grasping more meaningful knowledge from different feature levels is helpful to the precise segmentation of placental tissue.

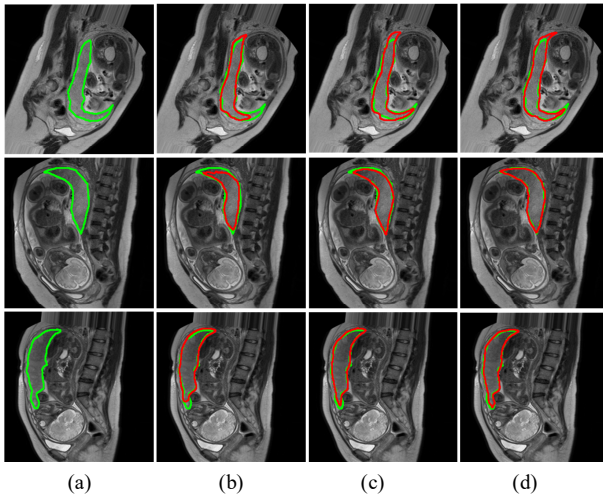
We further explore the effectiveness of each component of the context extractor module. The results are shown in Table 2. Both pyramid pooling module and transformer improve segmentation performance.

**Table 2 Ablation study on different components of context extractor module**

| Method                             | $M_{DSC}/\%$ |
|------------------------------------|--------------|
| TCGANet w/o transformer            | 85.21        |
| TCGANet w/o pyramid pooling module | 85.48        |
| TCGANet                            | 86.19        |

In Fig. 5, we visualized the segmentation results and found that our proposed model was likely to obtain more continuous segmentation boundary maps, and could completely segment the placental region. In Fig. 5(c), the result of segmentation without context extractor module is under segmentation, while in Fig. 5(d), the result of segmentation without discriminator is poor boundary segmentation. The experimental results show that our method can segment the irregular placental tissue well. These visualizations further demonstrate the importance of gathering multi-scale context information and capturing long-term dependencies in placental tissue segmentation.

To verify the effect of transformers of different sizes on the performance of placental tissue segmentation, this paper conducts ablation experiments on different transformer scales. The scale of the transformer is determined by the level ( $L$ ) of the transformer and the feature embedding dimension ( $d$ ). As shown in Table 3, when  $d=1024$  and  $L=12$ , the performance is the best.



**Fig. 5** Segmentation results of different methods (a) the original images layered with ground truth labels, (b)–(d) segmentation results, (b) without context extractor module, (c) without discriminator, (d) proposed TCGANet

**Table 3** Ablation study on transformer

| Level | Embedding dim | $M_{DSC}/\%$ |
|-------|---------------|--------------|
| 4     | 512           | 86.07        |
| 4     | 1 024         | 86.00        |
| 8     | 512           | 85.68        |
| 8     | 1 024         | 85.97        |
| 12    | 512           | 85.64        |
| 12    | 1 024         | 86.19        |

#### 4.3.2 Comparison of state-of-the-art methods

From the perspective of quantitative analysis, various methods are compared with different segmentation evaluation indicators. Each evaluation indicator uses the mean and standard deviation to summarize the results, as shown in Table 4. In Table 4, the various indicators of the method in this paper are better than the existing medical image segmentation

methods,  $M_{DSC}$  is  $0.903\pm0.093$ ,  $M_{SEN}$  is  $0.903\pm0.093$ , and  $M_{JSC}$  is  $0.778\pm0.169$ . The  $M_{JSC}$  of R2U-Net<sup>[10]</sup> is  $0.664\pm0.258$  slightly higher than that of U-Net<sup>[9]</sup>, and other indexes are similar to U-Net. The  $M_{SEN}$  and  $M_{DSC}$  of Att-Unet<sup>[11]</sup> are  $0.886\pm0.211$  and  $0.764\pm0.215$ , respectively, which are higher than those of U-Net, and it can be inferred that the attention-based network can improve the feature representation ability of placental tissue.

In addition, GAN, which uses U-Net as a generator, has achieved competitive performance in various indicators, the  $M_{PPV}$  and  $M_{DSC}$  are  $0.829\pm0.179$  and  $0.747\pm0.188$ , respectively, which are slightly lower than the method in this paper. The  $M_{PPV}$  of TransUNet<sup>[30]</sup> is higher than the above networks, reaching  $0.847\pm0.226$ . It shows that TransUNet can improve true positive by using the transformer for global modeling. The performance of our method in each metric was also compared with other methods using paired samples  $t$ -tests. The results indicated that most of the performance metrics of the proposed TCGANet were significantly better than the state-of-the-art method ( $P<0.05$ ).

The qualitative analysis is performed, and the differences between these existing medical image segmentation methods for placental tissue segmentation tasks are provided in detail, as shown in Fig. 6.

For the large placental tissue area, as shown in the first row of Fig. 6, the results of R2U-Net are only partial placental tissue pixels, which may be related to its feature extraction strategy. Att-Unet and TransUNet

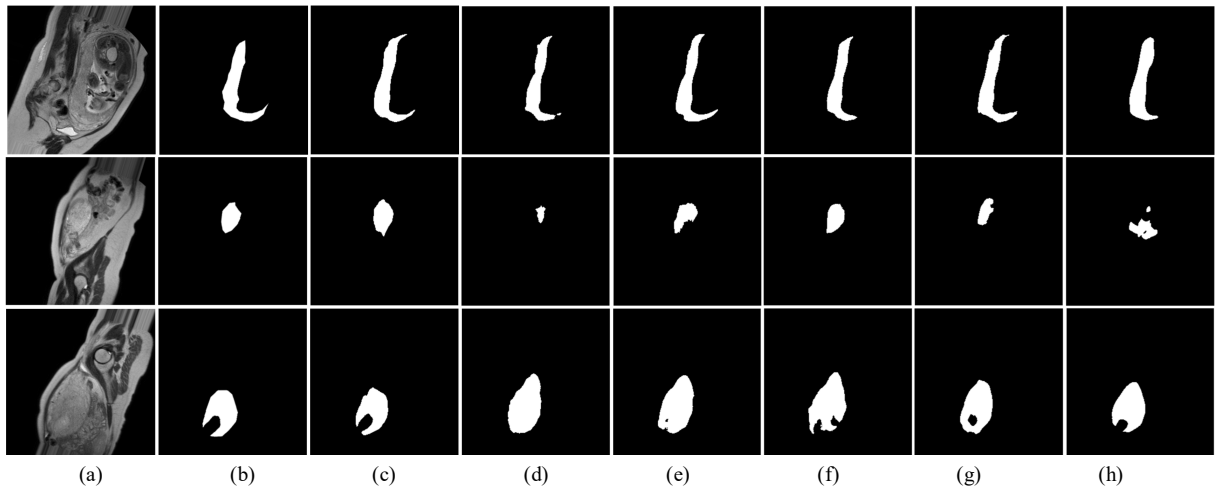
**Table 4** Quantitative comparison with different methods in placental tissue segmentation

| Method                   | $M_{ACC}$       | $M_{SEN}$       | $M_{SPE}$       | $M_{PPV}$       | $M_{DSC}$       | $M_{JSC}$       |
|--------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| U-Net                    | $0.991\pm0.005$ | $0.834\pm0.213$ | $0.993\pm0.005$ | $0.704\pm0.231$ | $0.748\pm0.229$ | $0.651\pm0.227$ |
| R2U-Net                  | $0.988\pm0.007$ | $0.838\pm0.273$ | $0.991\pm0.007$ | $0.664\pm0.258$ | $0.710\pm0.260$ | $0.664\pm0.258$ |
| Att-Unet                 | $0.990\pm0.005$ | $0.886\pm0.211$ | $0.992\pm0.006$ | $0.707\pm0.212$ | $0.764\pm0.215$ | $0.655\pm0.214$ |
| GAN(U-Net+discriminator) | $0.992\pm0.004$ | $0.882\pm0.135$ | $0.995\pm0.003$ | $0.829\pm0.179$ | $0.837\pm0.165$ | $0.747\pm0.188$ |
| TransUNet                | $0.987\pm0.006$ | $0.775\pm0.220$ | $0.973\pm0.005$ | $0.847\pm0.226$ | $0.809\pm0.223$ | $0.680\pm0.227$ |
| Our method               | $0.993\pm0.003$ | $0.903\pm0.093$ | $0.996\pm0.003$ | $0.848\pm0.171$ | $0.861\pm0.141$ | $0.778\pm0.169$ |

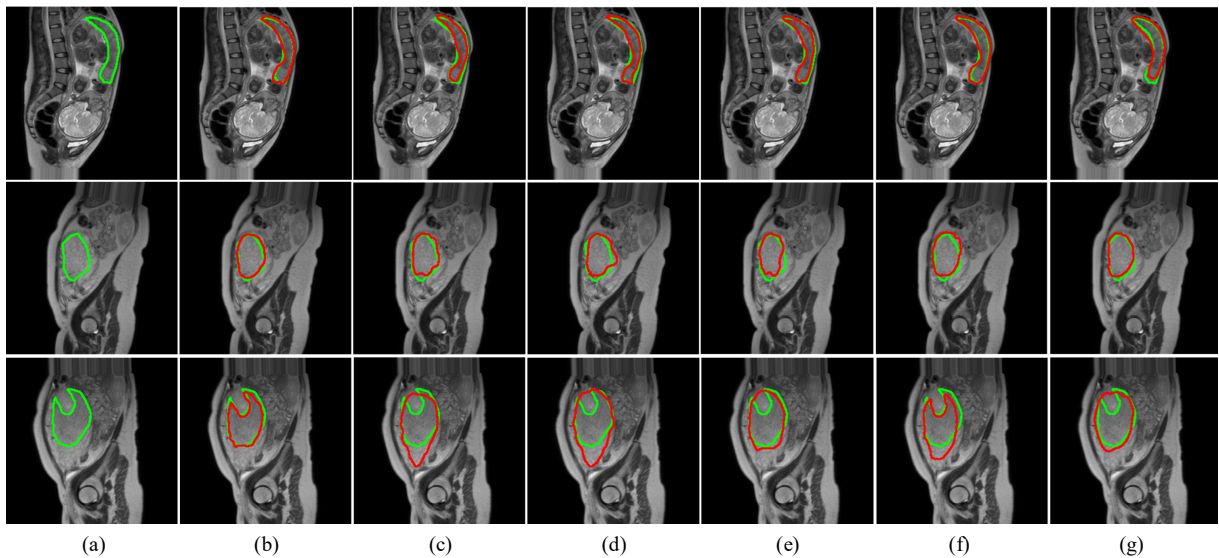
Note: Values are given as mean $\pm$ SD.

are over-segmentation, and U-Net and our method can predict the placental tissue more accurately. In the small placental tissue area shown in the second row of Fig. 6, U-Net, Att-Unet, GAN, and TransUNet did not completely predict the placental tissue. Only the segmentation result of our method is closest to the ground truth. It is also difficult for radiologists to distinguish these small placental tissues from other tissues on MR images, but the identification of these small placental tissues is also crucial for medical doctors

to detect the condition of the fetus. For the placental tissue with irregular regional shapes, as shown in the third row of Fig. 6, U-Net, Att-Unet, and R2U-Net all have over-segmentation. Over-segmentation will lead to higher false positives, which means that the shape of the placental tissue is missing in clinical applications. However, the method proposed in this paper can accurately segment the edge of the placental tissue. Fig. 7 shows a visual comparison of the accuracy of edge localization in placental tissue segmentation by different



**Fig. 6** Qualitative comparison with different methods in placental tissue segmentation, (a) the original images, (b) the ground truth, (c)–(h) segmentation results using (c) proposed TCGANet, (d) U-Net, (e) Att-Unet, (f) R2U-Net, (g) GAN(U-Net+discriminator), (h) TransUNet



**Fig. 7** Segmentation results of the proposed TCGANet and other competing models, (a) the original images layered with ground truth labels, (b)–(g) segmentation results using (b) proposed TCGANet, (c) U-Net, (d) Att-Unet, (e) R2U-Net, (f) GAN(U-Net+discriminator), (g) TransUNet

methods. As can be seen, our results are much closer to the ground truth. This shows that the proposed method is able to capture more dependencies between pixels and thus better capture the contours of the placental tissue.

## 5 Discussion

Automated placental tissue segmentation plays an important role in clinical diagnosis, as providing physicians with precise placental tissue localization and clear contour information can assist physicians in their diagnosis. Since placental tissue varies greatly in size in different slices and has a high similarity to surrounding tissue, there are still difficulties in segmenting placental tissue with large span differences and defining edges. Most of the existing medical segmentation models use an encoder-decoder structure, but this traditional skip connection method is prone to information loss. Generally, the shallow features contain more semantic information, such as the structure and boundary information of the placental tissue. On the contrary, the deep features contain more detailed information and are more capable to accurately describe the placental tissue. Therefore, these features should be fully exploited, and a feature embedding module is introduced in this paper.

Considering that the shortcoming of convolution is that it cannot capture long-distance pixel dependencies, it is not conducive to the segmentation of placental tissue with large spans. We designed the context extractor module, which uses the transformer to capture more global information to make up for the insufficiency of CNN and alleviate the limited receptive field of traditional segmentation networks. In addition, adversarial networks are used to supervise the generative network to improve the accuracy of placental edge definition.

Our proposed method still has some limitations. Fully supervised models rely on high-quality annotation data, so this method is suitable as an auxiliary tool for manual annotation, which can reduce the time of manual delineation, but cannot completely replace manual

segmentation.

## 6 Conclusion

In this paper, a deep learning model for placental tissue segmentation in MR images was proposed to help radiologists further medical image-based diagnosis, anatomical structure modelling and surgical planning.

To validate the effectiveness of the proposed method, various ablation experiments were conducted and the results of the comparison experiments showed that the method can achieve better performance in the task of placental tissue segmentation in MR images, outperforming existing medical segmentation networks. In future work, we will explore the extension of the proposed network to other tasks of medical image segmentation.

## References:

- [1] Jansen C H J R, Kastelein A W, Kleinrouweler C E, et al. Development of placental abnormalities in location and anatomy[J]. *Acta Obstetricia et Gynecologica Scandinavica*, 2020, 99(8):983-993.
- [2] Hecht J L, Baergen R, Ernst L M, et al. Classification and reporting guidelines for the pathology diagnosis of placenta accreta spectrum (PAS) disorders: Recommendations from an expert panel[J]. *Modern Pathology*, 2020, 33(12):2382-2396.
- [3] Sørensen A, Hutter J, Seed M, et al. T2\*-weighted placental MRI: Basic research tool or emerging clinical test for placental dysfunction?[J]. *Ultrasound in Obstetrics & Gynecology*, 2020, 55(3):293-302.
- [4] Maher M A, Abdelaziz A, Bazeed M F. Diagnostic accuracy of ultrasound and MRI in the prenatal diagnosis of placenta accreta[J]. *Acta Obstetricia et Gynecologica Scandinavica*, 2013, 92(9):1017-1022.
- [5] Chen Y, Hu F, Wang Y R, et al. Hybrid-attention densely connected U-Net with GAP for extracting livers from CT volumes[J]. *Medical Physics*, 2022, 49(2):1015-1033.
- [6] Bayramoglu N, Kannala J, Heikkilä J. Deep learning for magnification independent breast cancer histopathology image classification[C]//2016 IEEE 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 2017:2440-2445.

- [7] Xing J, Chen C, Lu Q Y, et al. Using BI-RADS stratifications as auxiliary information for breast masses classification in ultrasound images[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(6):2058-2070.
- [8] Zheng S Y, Guo J P, Cui X N, et al. Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection[J]. IEEE Transactions on Medical Imaging, 2020, 39(3): 797-805.
- [9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015:234-241.
- [10] Alom M Z, Yakopcic C, Hasan M, et al. Recurrent residual U-Net for medical image segmentation[J]. Journal of Medical Imaging, 2019, 6(1):014006.
- [11] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-Net: Learning where to look for the pancreas[EB/OL]. [2022-06-10]. <https://arxiv.org/abs/1804.03999>.
- [12] Araújo J D L, da Cruz L B, Ferreira J L, et al. An automatic method for segmentation of liver lesions in computed tomography images using deep neural networks[J]. Expert Systems with Applications, 2021, 180:115064.
- [13] McKinley R, Wepfer R, Aschwanden F, et al. Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks[J]. Scientific Reports, 2021, 11:1087.
- [14] Lei Y, Dong X, Tian Z, et al. CT prostate segmentation based on synthetic MRI-aided deep attention fully convolution network[J]. Medical Physics, 2020, 47(2): 530-540.
- [15] Valanarasu J M J, Oza P, Hacıhaliloglu I, et al. Medical transformer: Gated axial-attention for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, Cham, 2021:36-46.
- [16] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017:2881-2890.
- [17] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. [2022-06-10]. <https://arxiv.org/abs/1412.7062>.
- [18] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.
- [19] Li X M, Dou Q, Chen H, et al. 3D multi-scale FCN with random modality voxel dropout learning for inter-vertebral disc localization and segmentation from multi-modality MR images[J]. Medical Image Analysis, 2018, 45:41-54.
- [20] Li K Q, Qi X Q, Luo Y W, et al. Accurate retinal vessel segmentation in color fundus images via fully attention-based networks[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(6):2071-2081.
- [21] Shahedi M, Dormer J D, Anusha Devi T T, et al. Segmentation of uterus and placenta in MR images using a fully convolutional neural network[C]//SPIE Medical Imaging: Computer-Aided Diagnosis, Houston, Texas, USA, 2020, 11314:411-418.
- [22] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview[J]. IEEE Signal Processing Magazine, 2018, 35(1):53-65.
- [23] Luc P, Couprie C, Chintala S, et al. Semantic segmentation using adversarial networks[EB/OL]. [2022-06-10]. <https://arxiv.org/abs/1611.08408>.
- [24] Yang X Y, Zhang Y, Lo B, et al. DBAN: Adversarial network with multi-scale features for cardiac MRI segmentation[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(6):2018-2028.
- [25] Xue Y, Xu T, Zhang H, et al. SegAN: Adversarial network with multi-scale  $L_1$  loss for medical image segmentation [J]. Neuroinformatics, 2018, 16(3):383-392.
- [26] Dai W, Dong N Q, Wang Z Y, et al. SCAN: Structure correcting adversarial network for organ segmentation in chest X-rays[C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision, Springer, Cham, 2018:263-273.
- [27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. [2022-06-10]. <https://www.proceedings.com/content/034/034099webtoc.pdf>.
- [28] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[M]. Cham: Springer International Publishing, 2020:213-229.
- [29] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth  $16 \times 16$  words: Transformers for image recognition at scale[EB/OL]. [2022-06-10]. <https://arxiv.org/abs/2010.11929>.

- [30] Chen J N, Lu Y Y, Yu Q H, et al. TransUNet: Transformers make strong encoders for medical image segmentation [EB/OL]. [2022-06-10]. <https://arxiv.org/abs/2102.04306>.
- [31] Gao Y H, Zhou M, Metaxas D N. UTNet: A hybrid transformer architecture for medical image segmentation [M]//Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham: Springer International Publishing, 2021:61-71.
- [32] Zhang Y D, Liu H Y, Hu Q. TransFuse: Fusing transformers and CNNs for medical image segmentation [M]//Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham: Springer International Publishing, 2021:14-24.
- [33] Korkmaz Y, Dar S U H, Yurt M, et al. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers [J]. IEEE Transactions on Medical Imaging, 2022, 41(7): 1747-1763.
- [34] Dalmaz O, Yurt M, Cukur T. ResViT: Residual vision transformers for multimodal medical image synthesis[J]. IEEE Transactions on Medical Imaging, 2022, 41(10): 2598-2614.
- [35] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016:770-778.
- [36] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017:936-944.
- [37] Pang S C, Du A N, Orgun M A, et al. Tumor attention networks: Better feature selection, better tumor segmentation[J]. Neural Networks, 2021, 140:203-222.

## 基于结合 Transformer 和卷积神经网络的生成对抗网络 在磁共振成像中分割胎盘组织

叶正洁<sup>1</sup>, 王玉涛<sup>2</sup>, 徐建<sup>3</sup>, 金炜<sup>1\*</sup>

(1. 宁波大学 信息科学与工程学院, 浙江 宁波 315211; 2. 宁波大学医学院附属医院, 浙江 宁波 315020;  
3. 宁波妇女儿童医院, 浙江 宁波 315012)

**摘要:** 磁共振成像(MRI)胎盘组织的准确分割对于研究妊娠和分娩并发症具有重要意义, 但传统放射科医师的人工标注难以保证分割准确性和客观性, 且费时费力。为了开发用于 MRI 中胎盘组织自动分割的深度学习模型, 提出了结合 Transformer 和卷积神经网络(CNN)的生成对抗网络(TCGANet)。将特征嵌入模块与跳跃连接相结合, 缓解传统特征融合方法带来的信息丢失。在此基础上引入内容提取模块, 采用 Transformer 的自注意力机制捕捉全局依赖关系, 有效表示 MRI 的全局和局部信息。此外, 鉴于传统分割方法难于精确界定 MR 影像胎盘组织边缘的问题, 运用判别网络对胎盘组织分割的生成网络监督, 以提高胎盘边缘界定的精度。结果表明, 该模型在定量指标和边界定位精度方面显著优于现有分割方法, 其中准确度为  $0.993 \pm 0.003$ , 灵敏度为  $0.903 \pm 0.093$ , 特异度为  $0.996 \pm 0.003$ , Dice 相关性系数为  $0.861 \pm 0.141$ 。对模型不同结构的消融实验验证了网络结构设计的合理性, 大部分性能指标明显优于现有方法( $P < 0.05$ )。该模型能够实现自动且准确地分割 MRI 中胎盘组织。

**关键词:** 深度学习; 胎盘组织分割; 生成对抗网络; 核磁共振成像; 注意力机制

(责任编辑 史小丽)