Complex & Intelligent Systems (2023) 9:4503–4519
https://doi.org/10.1007/s40747-023-00968-x

ORIGINAL ARTICLE

DPC-MSGATNet: dual-path chain multi-scale gated axial-transformer network for four-chamber view segmentation in fetal echocardiography

Sibo Qiao[1] · Shanchen Pang[1] · Gang Luo[2] · Yi Sun[2] · Wenjing Yin[1] · Silin Pan[2] · Zhihan Lv[3]

Received: 3 September 2022 / Accepted: 3 January 2023 / Published online: 17 January 2023

Abstract

Echocardiography is essential in evaluating fetal cardiac anatomical structures and functions when clinicians conduct early treatment and screening for congenital heart defects, a common and intricate fetal malformation. Nevertheless, the prenatal detection rate of fetal CHD remains low since the peculiarities of fetal cardiac structures and the variousness of fetal CHD. Precisely segmenting four cardiac chambers can assist clinicians in analyzing cardiac morphology and further facilitate CHD diagnosis. Hence, we design a dual-path chain multi-scale gated axial-transformer network (DPC-MSGATNet) that simultaneously models global dependencies and local visual cues for fetal ultrasound (US) four-chamber (FC) views and further accurately segments four chambers. Our DPC-MSGATNet includes a global and a local branch that simultaneously operates on an entire FC view and image patches to learn multi-scale representations. We design a plug-and-play module, Interactive dual-path chain gated axial-transformer (IDPCGAT), to enhance the interactions between global and local branches. In IDPCGAT, the multi-scale representations from the two branches can complement each other, capturing the same region's salient features and suppressing feature responses to maintain only the activations associated with specific targets. Extensive experiments demonstrate that the DPC-MSGATNet exceeds seven state-of-the-art convolution- and transformer-based methods by a large margin in terms of F1 and IoU scores on our fetal FC view dataset, achieving a F1 score of 96.87% and an IoU score of 93.99%. The codes and datasets can be available at https://github.com\/QiaoSiBo/DPC-MSGATNet.

Keywords Convolutional neural network · Echocardiography · Fetal four chambers · Semantic segmentation · Transformer

Introduction

Congenital heart defect (CHD) is one of the most common inborn malformations, with the highest incidence of all congenital disability diseases and the leading cause of death in infancy [1]. Infants with CHD in China currently account for approximately 6‰–8‰ of all the born living neonates. Then, we can estimate that about 150,000 babies with CHD are born in China each year [2]. Therefore, the early diagnosis and recognition of CHD are of tremendous matter for the healthy growth of the fetus.

In recent years, echocardiography has been prevalently employed in clinical diagnosis and screening for pregnant women thanks to its quick imaging, low fees, and no radi-

✉ Shanchen Pang
pangsc@upc.edu.cn

✉ Silin Pan
silinpan@126.com

Sibo Qiao
siboqiao@126.com

Gang Luo
lg1989jay@163.com

Yi Sun
706666596@qq.com

Wenjing Yin
winsyin@126.com

Zhihan Lv
lvzhihan@gmail.com

[1] The School of Computer Science and Technology, China University of Petroleum, Qingdao 266580, Shandong, China

[2] The Heart Center, Qingdao Women and Children's Hospital, Qingdao 266034, Shandong, China

[3] The Department of Game Design Faculty of Arts, Uppsala University, Uppsala 75236, Sweden

ation exposure properties. In particular, echocardiography can effectively assess the fetal cardiac structure and function and plays a crucial position in recognizing and curing CHD [3]. The fetal ultrasound (US) four-chamber (FC) view provides clinicians with a clear view of the fetal cardiac morphology, the preferred view in prenatal diagnosis and examinations for fetal CHD [4]. In the early examinations of fetal CHD, the structural and functional parameters of the fetal heart are clinicians' primary object of evaluation [5]. It is worth mentioning that the segmentation of organs or lesions can quantitatively analyze the clinical parameters related to volume or developmental morphology, help clinicians accurately diagnose the patient's condition, and schedule a suitable treatment strategy [6].

For instance, the extraction of the ejection fraction of the left ventricle needs precise delineation of the left ventricular endocardium in both end-diastole and end-systole [7].

Motivation

Figure 1 shows the fetal FC structures, including the left atrium (LA), left ventricle (LV), right atrium (RA), and right ventricle (RV). Proper segmentation of fetal cardiac structures can provide an essential metric for evaluating fetal malformations. However, when analyzing FC views, the complex and variable structures of the fetal heart require clinicians to be professional in fetal cardiac anatomic structures and accurately measure parameters related to structure and function in a short period. Moreover, identifying and evaluating fetal cardiac structures and functions is a knowledge-intensive task that relies heavily on the extensive experience of clinicians. Therefore, it will be challenging for inexperienced clinicians to complete early diagnosis and examinations of fetal CHD. Simultaneously, the learning curve of this procedure may be very long due to several factors, such as the quality of the fetal ultrasound image, the different positions of the fetus in the womb, and the diversity of the fetal CHD [5]. As a result, a computer-aided system automatically segmenting the fetal four chambers will be highly welcomed to reduce the routine obstetric workload [8]. In addition, the computer-aided fetal cardiac segmentation system can also help medical novices learn through computerized feedback from score-based quality control procedures [9]. Furthermore, the computer-aided fetal cardiac segmentation system can provide pixel-level structural representations for other fetal FC view analysis tasks (e.g., classification), capture the pathological knowledge implied by ultrasound images, and further reduce empirical operations such as manual measurement of heart parameters. These operations can significantly improve the early diagnosis rate of fetal CHD.

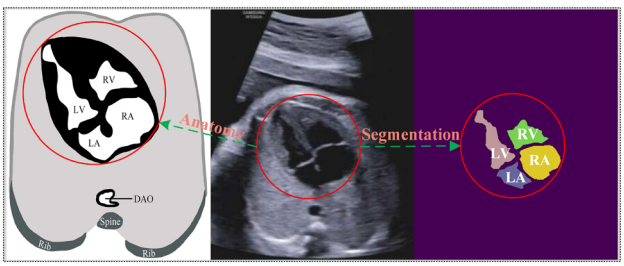

Fig. 1 The instance of the four chambers in a fetal FC view. The left column is fetal cardiac anatomical structures, the middle column is a fetal ultrasound FC view, and the right is fetal cardiac segmentation structures. As can be seen, the segmentation contours of the four chambers are very close to the anatomical contours

However, precise segmentation of fetal US FC structures faces the following challenges: first, the fetal US FC view often has poor image quality caused by diverse elements like imaging artifacts of acoustic shadows and speckles, deformation of soft tissues, fetal development, signal missing [9–11]. Second, the physical boundaries between the four chambers are not distinct or even disappear in the FC views when the mitral valve, tricuspid valve, atrium, or ventricle opens, making it more difficult to delineate the cardiac chambers accurately. Third, due to the involuntary movement of the fetus in the womb, position, or small heart size, there may be a high degree of similarity between FC structures in the FC view. Accordingly, even for experienced obstetricians, the category identification can be misled by the unique cardiac morphology [5,12]. Fourth, medical image data and expert annotations are significantly more limited and challenging to obtain than conventional computer vision tasks. Affected by the sonographers' technical level or the echocardiographic instrument's resolution, acquiring a large number of standard fetal FC views is a very time-consuming task. Meanwhile, labeling the fetal cardiac structures requires clinicians to process professional obstetric knowledge and is also time-consuming. In addition, due to the limited training data, the power of any machine learning-based computer-aided fetal FC segmentation method will be limited, making it challenging to obtain distinctive and robust representations to distinguish one identity from another. Hence, we should design a fetal FC segmentation model to capture context-invariant, position-sensitive, and identity-definite representations for fetal FC views.

Convolutional neural networks (CNNs) have achieved remarkable success in computer vision owing to their impressive feature learning ability, providing solid support for developing the computer-aided fetal FC segmentation system. Thanks to its inherent inductive bias in modeling local visual structures, CNNs can obtain excellent local features (e.g., edges and corners) by calculating local dependencies among neighbor pixels [13–15]. Moreover, rich low-level features are captured at shallow CNNs layers and then gradually aggregated into high-level semantic features through many stacked convolutional modules. Hence, many archi-

tectures based on CNNs have emerged in medical image segmentation [16–21]. These architectures achieve outstanding performance on various medical datasets, demonstrating the significance of CNNs in segmenting organs or lesions from medical images. Nevertheless, CNNs can only focus on local areas and cannot model global dependencies in an image. Moreover, long-distance dependencies are significant for medical image segmentation models, which should comprehend which pixels correspond to the targets and which pixels correspond to the background. Due to the background of an image being scattered, thus, capturing long-range dependencies between pixels corresponding to the background can help the model prevent background pixels from misclassifying as targets, thereby reducing false positives.

Transformer [22] has indicated a domination trend in almost all natural language processing benchmarks, attributed to their powerful ability to capture long-distance interactions among word tokens via the self-attention mechanism [23,24]. Subsequently, such excellent properties have inspired the development of traditional computer vision architectures [25–30]. In addition, several transformer-based approaches have been introduced to medical image segmentation recently [31–36], achieving more impressive performance than CNNs-based models. However, transformer-based methods usually require large-scale training data from these researches because (1) the applicable positional embedding required for a sequence of image tokens is challenging to learn from a small-scale dataset, and (2) they lack inherent inductive bias in modeling local visual structures and processing targets at different scales like convolutions.

Contribution

Inspired by the above viewpoints, we design two complementary strategies to solve the problem mentioned above: (1) we adopt the gated axial attention mechanism to control how much information is in positional embedding by applying four gates to key, query, and value parameters in self-attention [33] while factorizing 2D self-attention into two 1D self-attentions [37]. (2) We design a dual-path chain architecture that combines transformers with CNNs to model global and local dependencies to extract multi-scale representations for pixel-level dense segmentations. Hence, our contributions are mainly summarized as the following:

(1) We propose a **D**ual-**P**ath **C**hain **M**ulti-**S**cale **G**ated **A**xial-**T**ransformer **Net**work (DPC-MSGATNet) to segment four chambers from fetal US FC views. The DPC-MSGATNet includes a global and a local branch that simultaneously processes the entire FC views and image patches, capturing global and local visual cues to obtain multi-scale representations from fetal FC views.
(2) We propose an **I**nteractive **D**ual-**P**ath **C**hain **G**ated **A**xial-**T**ransformer (IDPCGAT) module to enhance the interactions between the global and local branches. The IDPCGAT is a plug-and-play module that captures the same region's salient features and suppresses feature responses to retain only the activations relevant to the specific targets.
(3) Our proposed DPC-MSGATNet outperforms seven state-of-the-art (SOTA) CNNs- and transformer-based methods by a large margin in terms of both F1 and IoU scores on the fetal US FC view dataset, achieving a F1 score of **96.87%** and an IoU score of **93.99%**.
(4) We adopt two public medical datasets to verify the generalization of the DPC-MSGATNet, which also achieves the best performance compared with the seven SOTA methods. Experimental outcomes show that

the DPC-MSGATNet acquires a F1 score of **85.22%** and an IoU score is **75.29%** on GLAS dataset [38] and a F1 score of **82.61%** and an IoU score of **70.69%** on MonuSeg dataset [39], respectively.

The rest of this paper is organized as follows: Sect. "Related work" provides several studies related to CNN-based and transformer-based segmentation methods in medical images. Then, we review several deep-learning methods in segmenting fetal cardiac anatomic structures from fetal FC views. Section "Our proposed DPC-MSGATNet" introduces our proposed DPC-MSGATNet and IDPCGAT module in detail. Section "Performance analysis" evaluates and discusses the performance of our proposed DPC-MSGATNet and its main components on the segmentation task. Finally, in Sect. "Conclusion", we present this paper's conclusion, our model's shortcomings, and future works.

Related work

This section summarizes the typical methods based on CNNs in medical image segmentation. Then, we review several of the transformer's related works in computer vision, particularly in medical image segmentation.

Medical image segmentation methods based on CNNs

CNNs are commonly used for image segmentation because of their powerful feature-learning capabilities. For example, for the first time, the fully convolution network (FCN) [40] abandons the full-connected layer in the model and uses full convolutions to semantically segment the image, directly demonstrating the feature expression ability of the CNNs. Further, the encoder-decoder-based U-Net [16], and its variants have shown excellent performance in medical

image segmentation. For instance, U-Net++ [17] designed a series of nested and dense skip connections to reduce the semantic gap between shallow and deep features. Attention U-Net [18] proposed a novel attention gate mechanism that automatically filters negative features from different levels in shortcut connections, allowing the model to focus on prominent features beneficial for segmentation targets. Res-UNet [19] added a weighted attention mechanism to the original U-Net [16], enabling the model to learn high-distinguished features that identify retinal blood vessels, thereby improving the performance of segmenting retinal vessels. DenseUNet [20] applied dense connections to the U-Net [16], allowing the model to explore the mixed representations of the liver and tumors end-to-end. UNet 3+ [21] employed full-scale skip connections and deep supervision to fuse high-level and low-level semantic features from different scales, further learning hierarchical representations from aggregated multi-scale features. Stacked U-Net [41] iteratively integrates features from various resolution scales while maintaining high spatial resolution at the output for recognizing small targets and sharp boundaries, enabling optimal segmentation performance with low computational complexity.

The above works all focus on improving network performance, but not too much attention is paid to computational complexity, inference time, or the number of parameters, which are crucial in many clinical diagnoses. Several networks [42,43], based on multilayer perceptron (MLP), have recently been proposed to be competent in computer vision tasks to reduce the computational overhead and accelerate the inference time. They can provide comparable performance to transformers yet with less computation. Furthermore, Valanarasu et al. [44] proposed a high-efficiency medical image segmentation model, UNeXt, which integrates CNNs and MLP to provide a more rapid inference time while keeping good performance; this makes it possible to deploy medical segmentation models in edge devices for rapid disease diagnosis.

In addition, methods based on CNNs have long been successfully applied to segment the cardiac anatomic structures, such as the LV [45,46], the RV [47,48], and biventricular segmentation [49]. Wang et al. [50] proposed a 2-stage improved U-Net model in which the RoI region of the heart is first automatically extracted in full-resolution cardiac CT and MR images, and then the whole heart is segmented into multiple categories in the RoI region. All of the above works are aimed at adult cardiac segmentation. Numerous studies have been conducted on segmenting cardiac anatomical structures in the fetal US FC views. Yu et al. [51] proposed a dynamic CNN model to segment the fetal LV in the fetal ultrasound images, which only selects a small LV area from the original ultrasound image for segmentation experiments. Xu et al. [8] proposed a cascading CNN model, DW-Net, which segments the LV, RV, LA, and RA in the fetal ultrasound FC views to be more consistent with clinical practice. Yang et al. [52] combined the data proportional balance strategy with Deeplab V3+ to segment the fetal ultrasound FC views.

Furthermore, several works [53–55] employ attention mechanisms to improve the segmentation performance. An et al. [53] proposed a category attention instance segmentation network (CA-ISNet) for the fetal four chambers segmentation. The CA-ISNet includes a category branch, a mask branch, and a category attention branch, which are used to predict the semantic category, segment the four chambers and extract category information of instances. Guo et al. [54] proposed a dual-path feature fusion network to segment LV and LA from fetal US FC views, which captures rich representations (e.g., high-level and low-level representations) via channel attention and spatial attention. Several works [56,57] adopt the advantages of the feature pyramid networks (FPN) [58] to extract multi-scale features, which is essential to capture high-level semantic and low-level boundary information. Pu et al. [57] proposed a MobileUNet-FPN to segment 13 anatomical structures in fetal US FC views, an encoder–decoder model combining the feature pyramid networks [58], and MobileNet [59]. To learn multi-scale features, Zhao et al. [60] proposed a two branches model, a multi-scale wavelet network (MS-Net), to segment LA and LV from FC views, which can capture detailed information through a discrete wavelet transform and bidirectional feature fusion. Table 1 summarizes several important segmentation works employing CNNs on cardiac anatomies. We can get those methods based on CNNs playing an essential role in medical image segmentation, especially the fetal cardiac anatomic structures.

Medical image segmentation methods based on transformer

Transformers are first applied to NLP and achieve excellent performance on machine translation. Moreover, it can compensate for several shortcomings of CNNs in capturing global context due to its ability to model long-range dependencies. Therefore, inspired by the success of transformers in various NLP, many researchers have explored its application to computer vision. For example, ViT [25] is a pioneering attempt to employ pure transformers, which requires large-scale datasets such as ImageNet-22K and JFT-300M to achieve SOTA performance in image classification. Furthermore, Swin Transformer [30] is a hierarchical transformer architecture, making the model have linear computational complexity by shifted windows strategy. Axial-DeepLab [37] decomposes the 2D self-attention into two 1D self-attentions to reduce the computational complexity and presents a position-sensitive axial attention scheme for segmentation. The transformer also shows outstanding

Table 1 Several important segmentation works employing CNNs on cardiac anatomic structures

Adult/fetus	Methods	Segmentation structures	Medical image type
Adult	DPM [45]	LV	MRI
	LDAMT [46]	LV	MRI
	CNNs and Stacked Autoencoders [47]	RV	MRI
	RegressionCNN [48]	RV	MRI
	SSLLN [49]	LV, RV	MRI
	Two-Stage U-Net [50]	LA, LV, RA, RV, AsAo, PA, MoLV	CT, MRI
Fetus	DW-Net [8]	LA, LV, RA, RV, TR, ED	US
	Dynamic CNN [51]	LV	US
	CA-ISNet [53]	LA, LV, RA, RV	US
	DP-PEM [54]	LA, LV	US
	AIDAN [55]	LA, LV	US
	MFP-UNet [56]	LV	US
	MobileUNet-FPN [57]	LA, LV, RA, RV, LVW, RVW, IS, LL, RL, SN, RB, DAO, IVS	US
	MS-Net [60]	LA, LV	US

performance on medical image segmentation. TransUNet [31], for example, uses transformers as a powerful encoder for U-Net, enhancing the detailed structures by restoring local spatial information when segmenting medical organs. TransFuse [32] blends transformers and CNNs in parallel, simultaneously learning global contextual information and low-level spatial detail. MedT [33] introduces gated axial attention from Axial-deeplab [37] and sets gating parameters to improve the accuracy of position embedding. At the same time, the global and local branches of MedT [33] learn different levels of image features from the whole image and the local image patches, respectively, to improve segmentation performance significantly. Karimi et al. [34] proposed a transformer deep neural network for 3D medical image segmentation, which splits 3D medical images into several 3D image patches and calculates the 1D embedding for each image patch. MBT-Net [35] fuses transformer and sketch structure branch to extract textured features and cell sketch position from corneal endothelial cell images. TransBTS [36] first uses a 3D CNN to extract brain MRI spatial feature mappings and then uses a transformer to model global dependencies for the extracted feature mappings. Inspired by the above methods, especially of MedT [33], we propose a dual-path chain encoder-decoder model based on transformer and CNNs to extract multi-scale local and long-range features and segment the fetal four chambers in FC views. Next, we will introduce our proposed model in detail.

Our proposed DPC-MSGATNet

Notations and problem definition

In this work, we adopt bold uppercase or lowercase letters (e.g., $\boldsymbol{X}, \boldsymbol{x}$) and uppercase letters (e.g., X) to represent matrics and scalars, respectively. For example, for a fetal US FC view, $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$, $\boldsymbol{X}$ is a matric that has three dimensions, the scalars of H, W, and C are the height, width, and the number of channels of $\boldsymbol{X}$. For an image patch, $\boldsymbol{x} \in \mathbb{R}^{\frac{H}{7} \times \frac{W}{7} \times C}$, $\boldsymbol{x}$ is also a matric that has three dimensions, and its height, width, and channel are defined by $\frac{H}{7}$, $\frac{W}{7}$, and C. Furthermore, we construct 49 patches from a fetal US FC view, that is $\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}]$, $N = 49$. The ground truth of the segmentation mask is described by $\boldsymbol{S} \in \mathbb{R}^{H \times W \times C}$, in which $C = 5$ represents the mask of LV, LA, RV, RA, and Background, respectively. The prediction of the segmentation mask is defined by $\hat{\boldsymbol{S}}$.

With the help of these notations, the purpose of this work is to jointly input $\boldsymbol{X}$ and $\boldsymbol{x_i}$ to capture discriminative representations to obtain expected segmentation mask $\hat{\boldsymbol{S}}$. From the prior knowledge, models based on the U-Net architecture can evenly distribute high- and low-resolution features from bottom-up across encoders and decoders, allowing the entire model to be trained end-to-end. During the decoding or deconvolution phase, the shallow high-resolution and high-level low-resolution feature maps are fused to produce an

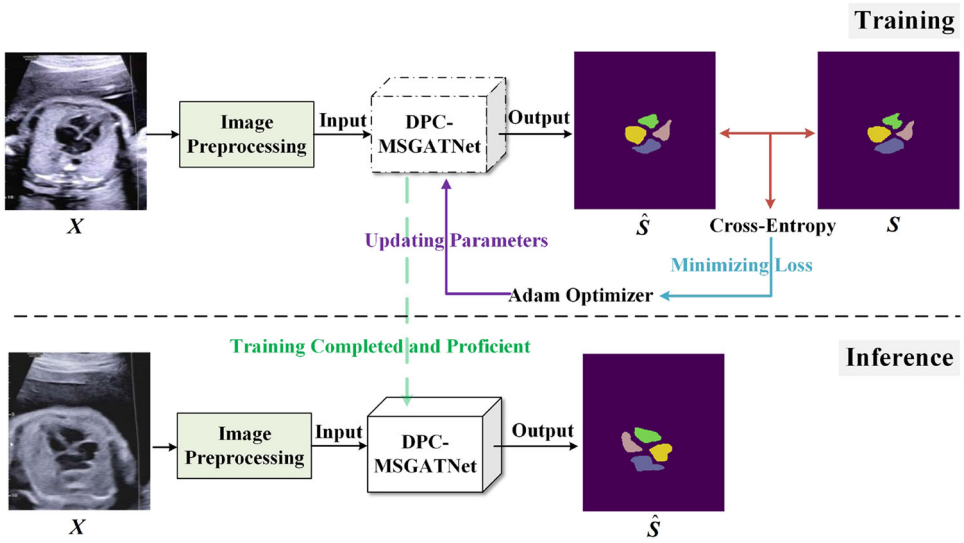

Fig. 2 The overall training and inference flowchart of DPC-MSGATNet for the fetal US FC view segmentation. In the training phase, given a fetal US FC view $\boldsymbol{X}$, we first adopt the image preprocessing methods to normalize and resize $\boldsymbol{X}$. Then, we feed it into the DPC-MSGATNet to obtain the predicted segmentation mask $\hat{\boldsymbol{S}}$. We employ Cross-Entropy to measure the distance between the predicted mask $\hat{\boldsymbol{S}}$ and ground truth $\boldsymbol{S}$. In order to quickly reduce the distance, we use the Adam to optimize the model and constantly update the model's parameters until the distance stabilizes. In the inference phase, we employ the trained DPC-MSGATNet, which already has specialized clinical knowledge for analyzing fetal US FC views, to obtain our desired segmentation of fetal four chambers

upsampled feature map by shortcut connections. However, these models do not perform well when faced with complex noise, artifacts, and low contrast in US images. Moreover, the high-resolution representations from the shallow layer fused by the decoder do not effectively encode rich semantic information. Hence, we propose a DPC-MSGATNet in this work, which is composed of two branches that can process bottom-up and top-down representations and capture long-distance interaction information at multiple scales. The general flowchart of DPC-MSGATNet with an example of fetal US FC view $\boldsymbol{X}$ as its input is illustrated in Fig. 2.

DPC-MSGATNet

As shown in Fig. 3, our proposed DPC-MSGATNet consists of the global branch that processes the whole image and the local branch that processes the image patches. These two branches comprehensively understand the input images at different scales, simultaneously capturing the image's high-level semantic information and long-distance spatial dependencies among image patches, further obtaining precise contours information of the segmented objects. However, transformer-based models are hungry for training data sets because of the required learning of appropriate position embedding, and medical images with high-quality annotations are expensive and challenging to collect. Therefore, the gated position-sensitive axial-transformer as the basic building module is adopted to encode input fetal US FC views in the two branches. In a gated position-sensitive axial-transformer, we adopt four gates to control how much information is learned by the positional embedding. These gates are all learnable parameters that enable the proposed network to be used for any data set of any size. Depending on the size of the training set, these gates will know if the number of the training set is enough to learn the proper positional embedding and then adaptively change depending on whether the information obtained by the position embedding is valid.

As shown in Fig. 3, we perform the operations for an input fetal US FC view $\boldsymbol{X}$ and image patch $\boldsymbol{x_i}$ in stage I, which is represented as follows:

$$\boldsymbol{C}_I = \mathcal{F}_{\text{global},I}(\boldsymbol{X}), \tag{1}$$

$$\boldsymbol{C}'_I = \mathcal{F}_{\text{local},I}(\boldsymbol{x}_i), \tag{2}$$

$$\boldsymbol{D}'_I = \Psi(\boldsymbol{C}'_I), \tag{3}$$

$$\boldsymbol{F}_{\text{fusion},I} = \boldsymbol{C}_I + \boldsymbol{D}'_I, \tag{4}$$

where $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ is an input FC view, C is the number of channels of an input FC view, $H \times W$ is the image size of each input FC view. $\boldsymbol{x}_i \in \mathbb{R}^{\frac{H}{7}\times\frac{W}{7}\times C}$ is a patch of an input FC view, $\frac{H}{7} \times \frac{W}{7}$ is the size of each image patch. $\mathcal{F}_{\text{global}}(\cdot)$ denotes a mapping of the global branch. $\mathcal{F}_{\text{local}}(\cdot)$ denotes a mapping of the local branch. $\Psi(\cdot)$ is the resample function for patches of an input FC view. $\boldsymbol{F}_{\text{fusion},I} \in \mathbb{R}^{H\times W\times C'}$ is the

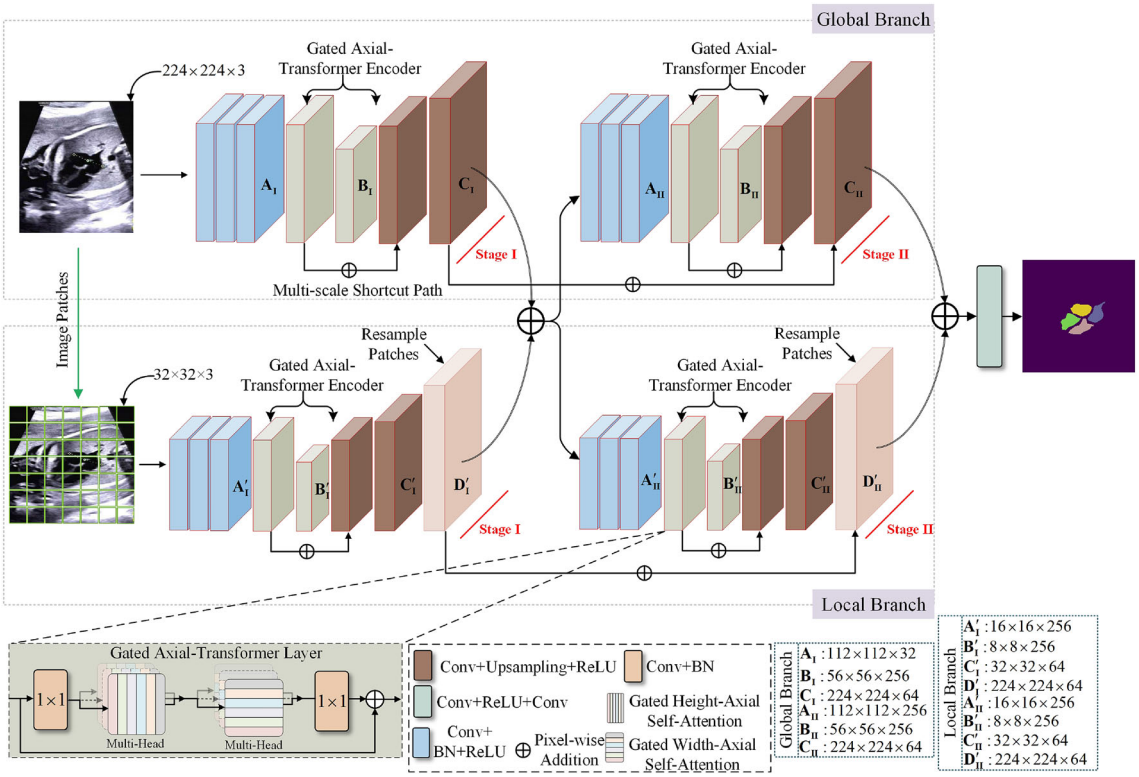

Fig. 3 The architecture of our proposed DPC-MSGATNet

fused feature map of global branch and local branch, $C^{'}$ is the number of channels of the fused feature map.

In stage II, we adopt the following operations to process the fused feature map $\boldsymbol{F}_{\text{fusion},I}$ and employ the shortcut connections to mix with the upsampled feature map in stage I:

$$\boldsymbol{C}_{II} = \mathcal{F}_{\text{global},II}(\boldsymbol{F}_{\text{fusion},I}) + \boldsymbol{C}_{I}, \tag{5}$$

$$\boldsymbol{C}'_{II} = \mathcal{F}_{\text{local},II}(\boldsymbol{f}_{\text{fusion},I}), \tag{6}$$

$$\boldsymbol{D}'_{II} = \Psi(\boldsymbol{C}'_{II}) + \boldsymbol{D}'_{I}, \tag{7}$$

$$\boldsymbol{F}_{\text{fusion},II} = \boldsymbol{C}_{II} + \boldsymbol{D}'_{II}, \tag{8}$$

where $\boldsymbol{f}_{\text{fusion},I} \in \mathbb{R}^{H\times W\times C^{'}}$ is a patch of the fused feature map $\boldsymbol{F}_{\text{fusion},I}$. Here, we define the computation in stage II as an interactive dual-path chain gated axial-transformer (IDPCGAT) module. As illustrated in Fig. 4, we can insert any number of the IDPCGAT modules depending on the size of the training set.

Then, we adopt the following operations to get the final segmentation:

$$\hat{\boldsymbol{S}} = \text{Conv}_{1\times1}(\sigma(\text{Conv}_{3\times3}(\boldsymbol{F}_{\text{fusion},II}))), \tag{9}$$

where $\hat{\boldsymbol{S}} \in \mathbb{R}^{H\times W\times C}$, $C = 5$. $\text{Conv}_{1\times1}(\cdot)$ represents a convolution operation with the filter size of 1×1, $\sigma(\cdot)$ represents a ReLU activation function, $\text{Conv}_{3\times3}(\cdot)$ represents a convolution operation with the filter size of 3×3. Next, we will introduce the main components of our DPC-MSGATNet in detail. More ablation studies on the architecture can be found in Tables 2, 3, and 4.

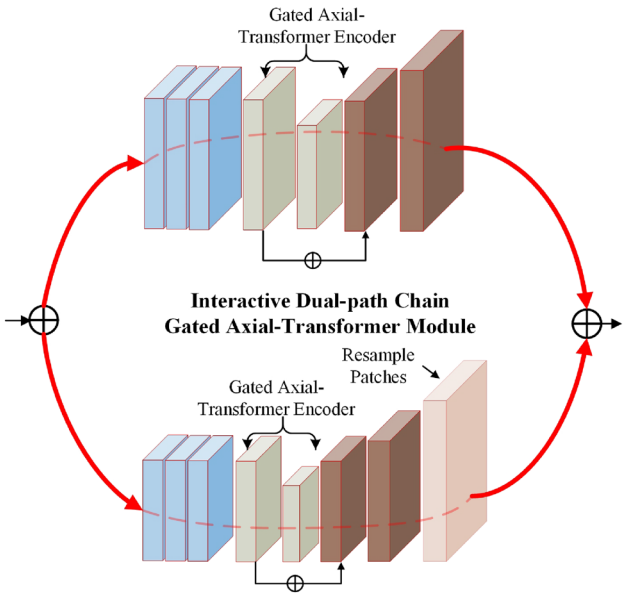

Fig. 4 The dual-path chain gated axial-transformer module

Table 2 Quantitative comparison against SOTA methods on the fetal FC view dataset

Methods	Fetal FC view dataset	
	F1 (%)	IoU (%)
Axial-Attention U-Net [37]	92.87 ± 0.48	86.98 ± 0.74
Gated-Axial-Attention U-Net [33]	93.94 ± 0.77	88.72 ± 1.23
U-Net [16]	95.40 ± 0.79	91.34 ± 1.27
Res-UNet [19]	95.47 ± 0.37	91.46 ± 0.60
U-Net++ [17]	95.55 ± 0.51	91.57 ± 0.83
Attention U-Net [18]	95.60 ± 0.51	91.66 ± 0.89
MedT [33]	95.69 ± 1.12	91.93 ± 1.91
DPC-MSGATNet	$\mathbf{96.87 \pm 0.73}$	$\mathbf{93.99 \pm 1.28}$

Bold metric means that its corresponding method performs best among other SOTA methods

Global/local branch

The global and local branches repeat the IDPCGAT module twice to achieve multi-scale fusion processing of features from bottom-up and top-down. When low-resolution shallow representations flow through multiple IDPCGAT modules, the upsampled high-resolution representations input to the decoder efficiently encode semantic information from the deep layers. The local branch is employed to capture more salient target details. Here we create 49 patches with the size of $\frac{H}{7} \times \frac{W}{7}$ in the local branch. Furthermore, each patch is fed forward through the local branch, and the output patch feature maps are resampled based on their relative location in the input FC views, obtaining the whole output feature maps. Hence, in stage I of the global and local branches, we first adopt three convolution operations to capture low-level representations of a fetal FC view $\boldsymbol{X}$, which can be represented as Eq. 10:

$$\begin{aligned} \boldsymbol{Z}_1 &= \sigma(\mathrm{BN}(\mathrm{Conv}_{7\times7}(\boldsymbol{X}))), \\ \boldsymbol{Z}_2 &= \sigma(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\boldsymbol{Z}_1))), \\ \boldsymbol{A}_I &= \sigma(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\boldsymbol{Z}_2))), \end{aligned} \tag{10}$$

where $BN(\cdot)$ represents Batch Normalization operation. For a patch $\boldsymbol{x}_i$, we also adopt Eq. 10 to learn low-level representations:

$$\boldsymbol{A}'_I = Eq.10(\boldsymbol{x}_i). \tag{11}$$

Then, the downsampled feature map $\boldsymbol{A}_I$ and $\boldsymbol{A}'_I$ are fed into a gated axial-transformer encoder, respectively. Here, a gated axial-transformer is represented as Eq. 12:

$$\begin{aligned} \boldsymbol{Z}_{i-1} &= \sigma(BN(Conv_{1\times1}(\boldsymbol{Z}_{i-1}))), \\ \boldsymbol{Z}_i &= MHHASA(\boldsymbol{Z}_{i-1}), \\ \boldsymbol{Z}_{i+1} &= MHWASA(\boldsymbol{Z}_i), \\ \boldsymbol{Z}_{i+2} &= \sigma(BN(Conv_{1\times1}(\boldsymbol{Z}_{i+1}))), \end{aligned} \tag{12}$$

Table 3 Quantitative comparison of the branch ablation study in our DPC-MSGATNet

Methods	Fetal FC view dataset	
	F1 (%)	IoU (%)
Local Branch	87.54 ± 5.01	77.98 ± 5.96
Global Branch	96.63 ± 0.32	93.61 ± 0.57
DPC-MSGATNet	$\mathbf{96.87 \pm 0.73}$	$\mathbf{93.99 \pm 1.28}$

Bold metric means that its corresponding method performs best among other SOTA methods

Table 4 Quantitative comparison of the CI ablation study in our DPC-MSGATNet

Methods	Fetal FC view dataset	
	F1 (%)	IoU (%)
DPC-MSGATNet without CI	96.59 ± 0.88	93.50 ± 1.55
DPC-MSGATNet	$\mathbf{96.87 \pm 0.73}$	$\mathbf{93.99 \pm 1.28}$

Bold metric means that its corresponding method performs best among other SOTA methods

where MHHASA$(\cdot)$ denotes a multi-head height-axial self-attention operation. MHWASA$(\cdot)$ denotes a multi-head width-axial self-attention operation. We adopt GAT$(\cdot)$ to denote Eq. 12. Hence, a gated axial-transformer encoder for the $\boldsymbol{A}_I$ is represented as Eq. 13:

$$\begin{aligned} \boldsymbol{Z}_3 &= GAT(\boldsymbol{A}_I), \\ \boldsymbol{Z}_4 &= GAT(\boldsymbol{Z}_3), \\ \boldsymbol{B}_I &= GAT(\boldsymbol{Z}_4). \end{aligned} \tag{13}$$

Here we also adopt Eq. 13 to encode $\boldsymbol{A}'_I$:

$$\boldsymbol{B}'_I = Eq.13(\boldsymbol{A}'_I). \tag{14}$$

Then, we begin to decode the $\boldsymbol{B}_I$ by Eq. 15:

$$\begin{aligned} \boldsymbol{Z}_5 &= \sigma(Upsampling(Conv_{3\times3}(\boldsymbol{B}_I))), \\ \boldsymbol{C}_I &= \sigma(Upsampling(Conv_{3\times3}(\boldsymbol{Z}_5))), \end{aligned} \tag{15}$$

where Upsampling(·) is a bilinear interpolation method in this work. Here we also employ Eq. 15 to decode $\boldsymbol{B}'_I$, and then to resample decoded patches:

$$\boldsymbol{C}'_I = \text{Eq. } 15(\boldsymbol{B}'_I), \tag{16}$$

$$\boldsymbol{D}'_I = \Psi(\boldsymbol{C}'_I). \tag{17}$$

Gated axial-transformer

To an input medical image $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$, we flatten it into a matrix $\boldsymbol{X} \in \mathbb{R}^{HW\times C}$ and conduct multi-head self-attention operation as proposed in the transformer [22]. Moreover, the output of the self-attention module for a single head can be formulated as follows:

$$\boldsymbol{O} = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{18}$$

where queries $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q$, keys $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k$, values $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_v$. $\boldsymbol{W}_q$, $\boldsymbol{W}_k \in \mathbb{R}^{C\times d_k}$ and $\boldsymbol{W}_v \in \mathbb{R}^{C\times d_v}$ are learnable projection parameter matrices for the input $\boldsymbol{X}$. Hence, the outputs of all heads are computed as follows:

$$\text{MHSA}(\boldsymbol{X}) = \text{Concat}[\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots \boldsymbol{O}_h]\boldsymbol{W}^O, \tag{19}$$

where $\boldsymbol{W}^O \in \mathbb{R}^{d_v\times d_v}$ is a learnable projection parameter matrix. The self-attention mechanism allows transformers to model long-distance dependencies between pixel tokens or captures non-local information from the whole feature map. That is why transformers have had excellent success in language and vision. However, transformer networks are hungry for data sets to achieve state-of-the-art performance. Therefore, as shown in Fig. 5, limits to the medical image scale, we adopt a gated axial-transformer as proposed in Axial-DeepLab [37] to perform self-attention on the height axis and width axis of the feature map, respectively. Hence, for instance, a gated axial self-attention for the height axis in the gated axial-transformer layer is formulated as:

$$\begin{aligned} O_{ij} = \sum_{h=1}^{H} \text{softmax}(Q_{ij}K_{ih}^T + G_q Q_{ij} R_{ih;q}^T \\ + G_k K_{ih} R_{ih;k}^T)(G_{v_1} V_{ih} + G_{v_2} R_{ih;v}), \end{aligned} \tag{20}$$

where G_q, G_k, G_{v_1} and G_{v_2} are all learnable gating parameters. $R \in \mathbb{R}^{H\times H}$ is relative positional encoding. Initially initialized to 1.0, gating parameters are employed to control the influence of the relative positional encodings in a non-local context.

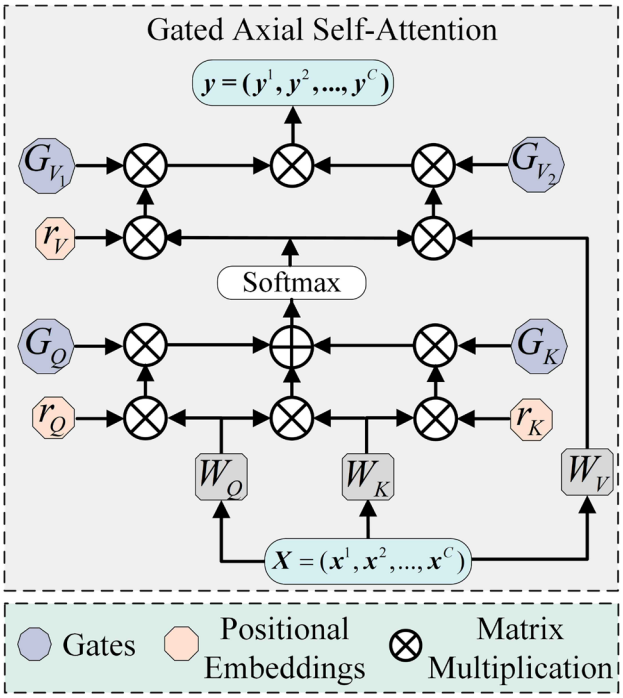

Fig. 5 The architecture of the gated axial self-attention

Performance analysis

Datasets and preprocessing

We obtain the fetal FC view dataset from the Qingdao Women and Children's Hospital. We randomly selected 556 FC views of fetuses from the hospital from 24 to 26 weeks of gestation. These views are collected from 600 fetuses as the entire experimental dataset, which has different degrees of artifacts, speckle noise, and inconspicuous borders, making them very convenient for confirming the effectiveness of the DPC-MSGATNet in haggling with the segmentation task. Furthermore, two professional radiologists tag all the views employed in this work, and then the annotated views undergo rigorous verification. In addition, we randomly split the entire experimental dataset. Here the training set consists of 446 FC views, and the test set includes 110 FC views that do not appear in the training set. The original experimental dataset indicates that each FC view has a different size. Hence, we resized the FC view into 224 × 224. In addition, the fetal FC view and its corresponding mask are augmented by random horizontal and vertical flip operations to relieve over-fitting.

In addition, two public medical datasets, Gland Segmentation (GLAS) [38] and MonuSeg [39], are also used to evaluate our DPC-MSGATNet. GLAS contains 165 microscopic images and their related ground-truth mask. Here we split it into 132 images for training and 33 for testing. Due to images in the GLAS having different scales, in our exper-

iments, we resize each image to a resolution of 224 × 224. The MonuSeg contains 46 tissue images and the corresponding ground-truth mask. Here we split it into 37 images for training and 9 for testing. Finally, we resize per image into 128 × 128.

Implementation details

Hyper-parameters setting. In our experiments, the training step is 400 epochs, and the mini-batch size is 4. Furthermore, the initial learning rate is 0.001. We employ the Adam optimizer to optimize the DPC-MSGATNet, whose weight decay is 0.00001. The ReduceLROnPlateau strategy is used to adjust our initial learning rate when the loss is not changing, wherein the factor and patience are set at 0.8 and 15, respectively. We do not train the four gates for the first 50 epochs when training the gated axial attention layer. We randomly divide the dataset 5 times and perform a 5-fold cross-validation. For more detailed parameter settings of the DPC-MSGATNet, please refer to the codes provided in this article.

Hardware setting. We have a GPU cluster that mainly includes a management node, a GPU node, a storage node, and a storage array. The management node is a DELL EMC PowerEdge R740 server with one physical CPU, and its version is 4214 (12 CPU cores and 24 logical processors). The GPU node has two DELL EMC PowerEdge R740 servers, each with two physical CPUs, and its version is 6226R (16 CPU cores and 32 logical processors). Each GPU server has two NVIDIA Telsa V100 32G. The storage node is a DELL EMC PowerEdge R740 server with two physical CPUs, and its version is 4216 (16 CPU cores and 32 logical processors). The storage array is a DELL EMC ME4024 server with a 40TB HDD.

In this work, all the transformer-based models are trained with two NVIDIA Tesla V100 32G in a parallel computing manner, and all the CNN-based models are trained with one NVIDIA Tesla V100 32G. We adopt an NVIDIA 3090 24G workstation to conduct inference tests when these models finish training. The NVIDIA 3090 24G workstation equips one Inter i7-10700 CPU with 8 CPU cores and 16 logical processors.

Software setting. In this work, we employ Python 3. 7. 6, Pytorch 1. 7. 1, and Torchvision 0. 8. 2 to implement our DPC-MSGATNet. Furthermore, in the GPU environment, CUDA and CUDNN are 11. 0. 221 and 8. 0. 5, respectively. In Python 3. 7. 6, the Numpy is 1. 20. 1, the Opencv is 3. 4. 2, and the Pillow is 8. 2. 0.

Objective function. To finely measure the dissimilarity between our predicted segmentation mask and the ground truth, we adopt Cross-Entropy as our loss function to optimize our proposed model. Here the loss function used in this work is computed by:

$$\text{Loss} = -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N} g_i^c \log p_i^c, \tag{21}$$

where g_i^c is the ground truth binary indicator of class c of pixel i, and p_i^c is the corresponding predicted segmentation probability.

Evaluation measures

To evaluate the segmentation performance of our proposed DPC-MSGATNet, we adopt two general methods, F1 and IoU scores, to measure the similarity between the ground-truth mask and the predicted segmentation mask. The F1 and IoU are computed by:

$$F1 = \frac{2N_{\text{TP}}}{2N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}}}, \tag{22}$$

$$\text{IoU} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}}}, \tag{23}$$

where N_{TP} represents the number of pixels marked with the class 1 and predicted by the DPC-MSGATNet to be 1, N_{FP} represents the number of pixels marked with the class 0, yet predicted to be 1. N_{FN} represents the number of pixels marked with class 1 yet predicted to be 0.

Results and discussion

In this subsection, we begin by comparing the segmentation performance of our DPC-MSGATNet against the current SOTA methods. Then, we perform detailed ablation studies to analyze the contributions of components of our DPC-MSGATNet and conduct an inference time comparison with the SOTA methods. Finally, two public medical datasets are adopted to demonstrate the generalization performance of our DPC-MSGATNet.

Comparison with SOTA methods. To make a fair performance comparison, we choose several SOTA methods based on CNN and transformer, which include U-Net [16], U-Net ++ [17], Attention U-Net [18], Res-UNet [19], Axial-Attention U-Net [37], Gated-Axial-Attention U-Net [33], and MedT [33], respectively. The above two metrics, F1 score, and IoU are used to evaluate the segmentation performance of these models. Furthermore, U-Net [16], U-Net++ [17], Attention U-Net [18], and Res-UNet [19] are all based on CNN. Axial-Attention U-Net [37], Gated-Axial-Attention U-Net [33], MedT [33], and our DPC-MSGATNet are all transformer-based attention models. Table 2 shows the quantitative comparison of our DPC-MSGATNet with the two

kinds mentioned above of methods. From Table 2, our proposed DPC-MSGATNet achieves the best performance among all the SOTA methods, in which the F1 score and IoU are 96.87% and 93.99%, respectively. On the other hand, the Axial-Attention U-Net [37] has the worst performance, and the F1 score and IoU are 92.87% and 86.98%, separately.

From Table 2, except for the Axial-Attention U-Net [37] and Gated-Axial-Attention U-Net [33], the transformer-based attention models achieve better segmentation performance on the FC views dataset than the CNN-based models. The Attention U-Net [18] realizes the best effect with the F1 score of 95.60% and IoU of 91.66% among these CNN-based models. Therefore, our DPC-MSGATNet improves by 1.27% and 2.33% than the best convolutional baseline in terms of F1 score and IoU. For the transformer-based attention models, our proposed DPC-MSGATNet outperforms MedT [33] by a large margin in terms of both F1 score and IoU, which also processes global and local branches to capture fetal four-chamber multi-scale non-local content. The F1 score and IoU are improved by 1.18% and 2.06%, respectively. Transformer-based models require lots of training data to learn SOTA representations because of the positional embedding. It is worth mentioning that the Gated-Axial-Attention U-Net [37] performs better than the Axial-Attention U-Net [33], demonstrating that the gated mechanism works in controlling the information learned by the positional embedding.

Figure 6 illustrates the visual segmentation results of fetal FC views by DPC-MSGATNet and 7 SOTA methods. As shown in Fig. 6, our DPC-MSGATNet performs best on the segmentation of fetal FC views, in which the FC contours predicted by DPC-MSGATNet are closest to the ground truth. Furthermore, we can notice that the CNN-based models are prone to misclassification. For example, in the fourth row of Fig. 6, the segmentation mask predicted by the U-Net [16], U-Net++ [17], Attention U-Net [18], and Res-UNet [19] shows that more background pixels are incorrectly labeled as positives. On the contrary, except for Axial-Attention U-Net [37] and Gated-Axial-Attention U-Net [33], the transformer-based attention models such as MedT [33] and our DPC-MSGATNet, precisely identify which pixels correspond to the positives and which to the background. The Axial-Attention U-Net [37] and Gated-Axial-Attention U-Net [33] are inferior to the CNN-based models.

Unfortunately, transformers require large-scale training data to learn excellent positional encodings. Nevertheless, this is a dilemma for medical images because collecting and labeling large-scale medical datasets is very time-consuming and expensive. In this work, the fetal FC views training data is limited. The Axial-Attention U-Net [37] is based on traditional transformers, which require large-scale training data and a more extended training schedule. The Gated-Axial-Attention U-Net [33] adopts gated parameters to control the amount of information obtained by the positional embedding, thereby achieving better performance than Axial-Attention U-Net [37] by reducing the dependencies on the number of the training dataset. Moreover, the MedT [33] and DPC-MSGATNet design a unique architecture that includes a global and local branch to capture long-range interactions among image patches through global, learnable, and adapted attention coefficients to the input images. In particular, our DPC-MSGATNet performs this operation better. We propose a chain architecture in DPC-MSGATNet, enhancing the interactions between the global and local branches. Furthermore, our extensive experiments found that when to train four gating parameters is also critical to the segmentation performance of the model. If the four gating parameters are trained from the beginning, the model may be unstable or turbulent during training due to the relatively scarce training data, leading to a decrease in the model's performance. Therefore, in this work, we employ a training trick in that the four gating parameters are initialized with 1.0 and trained after 50 epochs. As we all know, transformers do well in modeling global dependency using the self-attention mechanism, yet they lack an intrinsic inductive bias in extracting local visual context. Our DPC-MSGATNet combines convolutions with transformers to learn abundant multi-scale representations of FC views. The above analysis indicates why our DPC-MSGATNet outperforms 7 SOTA models in segmenting fetal four chambers.

Ablation study. To analyze the contributions of each component in our DPC-MSGATNet, we perform detailed ablation studies on the branch, chain interactions (CI), and layers. All the models are trained for 400 epochs on the fetal FC views dataset and follow the same training strategy described in Sects. "Datasets and preprocessing" and "Implementation details". Next, we will conduct a detailed discussion on the ablation study.

Branch ablation. As shown in Table 3, we investigate the sub-structures in our DPC-MSGATNet, namely local and global branches, by isolating them separately. For example, an input image of size 224×224 is first split into 49 image patches of 32×32 in the local branch. Then, these image patches are continually fed into the local branch to extract local representations. As can be seen, the local branch performs poorly than the global branch, only achieving a F1 score of 87.54% and an IoU score of 77.98%. On the other hand, the global branch performs better, in which the F1 score and IoU are improved by 9.09% and 15.63%, separately.

Figure 7 shows the visual segmentation results of different structures in our DPC-MSGATNet. We also observed that the global branch performs much better than the local branch, close to the DPC-MSGATNet's performance. The global branch is fed into a whole image, which is essential in improving the performance of DPC-MSGATNet. On

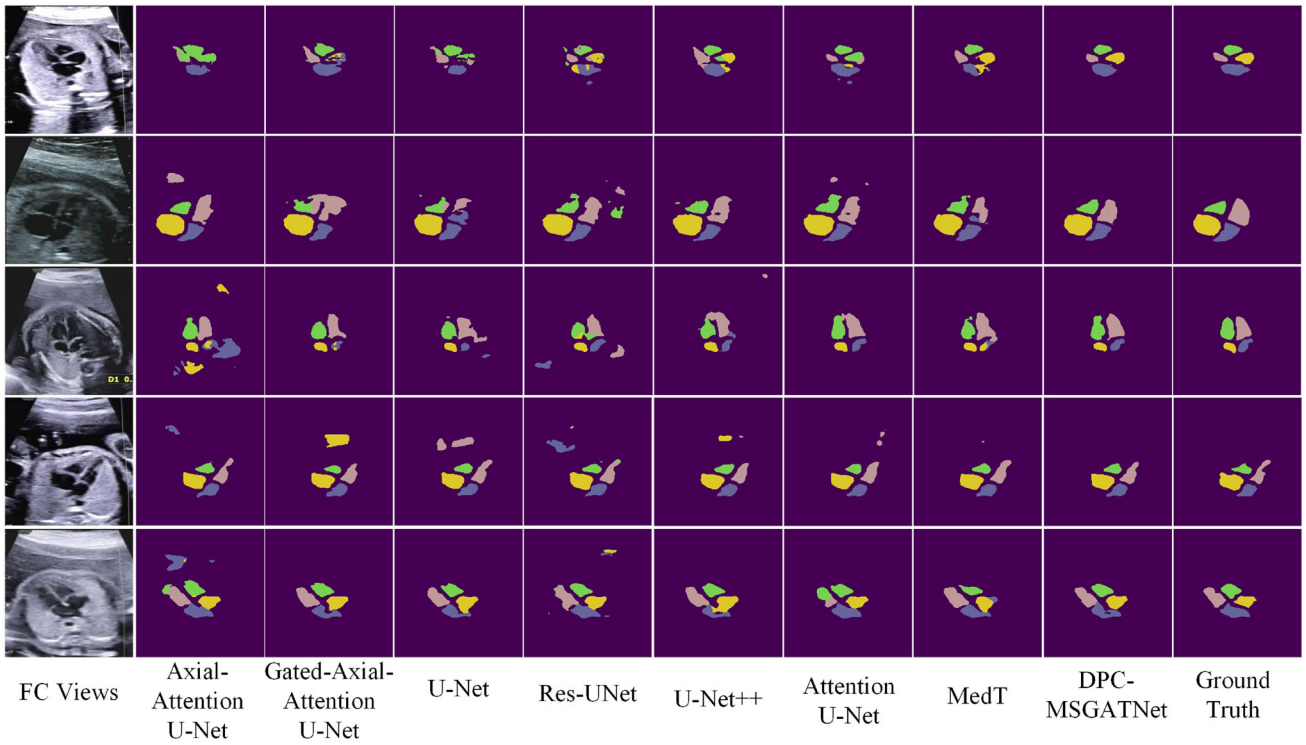

Fig. 6 Visual segmentation comparison of fetal FC views by DPC-MSGATNet with 7 SOTA methods

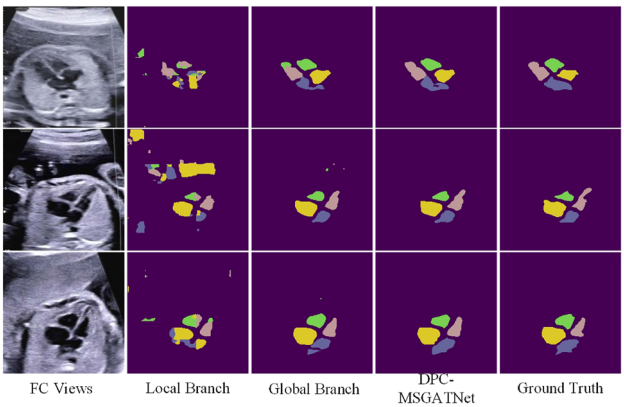

Fig. 7 Visual segmentation comparison of the branch ablation study in our DPC-MSGATNet

the other hand, the local branch is fed into image patches, which are limited to focusing on the local physical area of the input image, fail to establish a good connection with other image patches, and easily ignore the whole image's contextual correlation information. Nevertheless, the local branch can provide more detailed contours of four chambers which the global branch ignored. Furthermore, the chain structures in our DPC-MSGATNet can increase the interactions between global and local branches. Then, we can obtain more comprehensive representations that encode both the global context and local visual cues.

Hence, DPC-MSGATNet outperforms any sub-architectures, improving the segmentation performance on the fetal FC views dataset by 0.24% in F1 score and 0.38% in IoU score than the global branch.

CI ablation. To demonstrate the effectiveness of the proposed IDPCGAT in fusing multi-scale representations, we compare our DPC-MSGATNet with DPC-MSGATNet without CI in the same training settings. DPC-MSGATNet without CI means no interactions between the global and local branches, and the representations learned by the two branches are not fused until the end of the model.

The quantitative results are shown in Table 4. As can be seen, DPC-MSGATNet without CI also achieves good performance on the segmentation task, in which the F1 score and IoU are 96.59% and 93.50%, respectively. The interactive integrations between multi-scale representations are significant, and thereby the IDPCGAT helps our DPC-MSGATNet achieve better performance, in which the F1 score and IoU are enhanced by 0.28% and 0.49%, respectively. It is worth noting that the global branch performs better than the DPC-MSGATNet without CI, improving by 0.04% and 0.11% in terms of F1 score and IoU, respectively. Although the local branch can bring more delicate representations to the whole DPC-MSGATNet, if the model does not fuse and absorb advantages from the global branch throughout the training process but only perform a straight fusion at the end of the

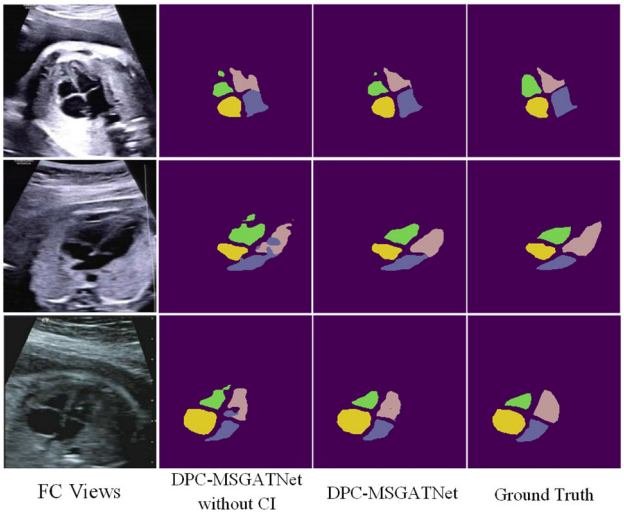

Fig. 8 Visual segmentation comparison of the IDPCGAT ablation study in our DPC-MSGATNet

model, it is easy to confuse the learned representations, which in turn makes the model have poor performance.

Figure 8 shows the visual segmentation effect of the IDPCGAT in our DPC-MSGATNet. It can be seen that the IDPCGAT makes the model automatically identify salient feature map regions and fuse feature responses from multi-scale representations to conserve only the activations relevant to the fetal four chambers. Patches of various scales can complement each other in representation extraction. Large patches can better capture coarse-grained representations, while small patches can better capture fine-grained representations. Hence, increasing the interactions between multi-scale representations can improve performance on segmentation.

Layers ablation. As shown in Fig. 3, we stack two IDPCGAT modules in our DPC-MSGATNet. Then, we adopt a shortcut path to transfer high-resolution feature maps between the two IDPCGAT modules. As the data flows through multiple IDPCGAT modules, high-resolution feature maps from shallow layers also can encode rich semantic context information. In this work, we build our base model, DPC-MSGATNet-S, by stacking two IDPCGAT modules. In addition, we also introduce DPC-MSGATNet-T, DPC-MSGATNet-B, and a giant version of DPC-MSGATNet-L, which are about 0.34×, 1.65×, and 2.32× model parameters, respectively. The architectures of these models are as follows:

- DPC-MSGATNet-T: stacking IDPCGAT numbers = 1
- DPC-MSGATNet-S: stacking IDPCGAT numbers = 2
- DPC-MSGATNet-B: stacking IDPCGAT numbers = 3
- DPC-MSGATNet-L: stacking IDPCGAT numbers = 4

The chain interaction, IDPCGAT, is a plug-and-play module. When we have a large-scale training dataset, we can stack more IDPCGAT with no bells and whistles. Table 5 and Fig. 9 show quantitative results and visual segmentations. As can be seen, our DPC-MSGATNet outperforms three variants on the fetal FC views in this work. Furthermore, the performance of the DPC-MSGATNet-B outperforms DPC-MSGATNet-T and DPC-MSGATNet-L. We suspect this phenomenon may be related to the scale of the dataset, and more data will be collected in the future to validate the conjecture.

Inference time. In clinical practice, clinicians often want to be able to detect diseases effectively and give reasonable treatment measures in the shortest possible time. Therefore, their inference speed is critical for clinical diagnosis when computer-aided models are deployed on edge devices or AI servers. We conduct an inference time test on the test dataset in this work. Table 6 compares SOTA methods' inference time mean on NVIDIA GPU 3090. Table 6 shows that CNN-based methods have less inference time for fetal FC US views segmentation than transformer-based attention methods, in which the U-Net has a minimum inference time during all the CNN-based methods. Our DPC-MSGATNet has the maximum inference time compared with the 7 SOTA methods in the manuscript yet has the best segmentation performance. For the segmentation task of a fetal US FC view, the inference time of DPC-MSGATNet is 0.8464 seconds. It is worth mentioning that the Global Branch of our DPC-MSGATNet has a lower inference time, 0.4869 seconds, which is not much different from the inference time of CNN-based methods.

However, the Global Branch has a better segmentation performance than the 7 SOTA methods, only 0.24% less than the DPC-MSGATNet in F1 score.

Generalization on two public medical datasets. To further analyze the generalization of our proposed DPC-MSGATNet

Table 5 Quantitative comparison of the layers ablation study in our DPC-MSGATNet

Methods	Params (M)	Fetal FC view dataset	
		F1 (%)	IoU (%)
DPC-MSGATNet-T	**2.2**	96.46 ± 0.74	93.25 ± 1.29
DPC-MSGATNet-B	10.8	96.62 ± 0.85	93.54 ± 1.45
DPC-MSGATNet-L	15.2	96.27 ± 0.59	92.92 ± 1.04
DPC-MSGATNet-S	6.5	**96.87 ± 0.73**	**93.99 ± 1.28**

Bold metric means that its corresponding method performs best among other SOTA methods

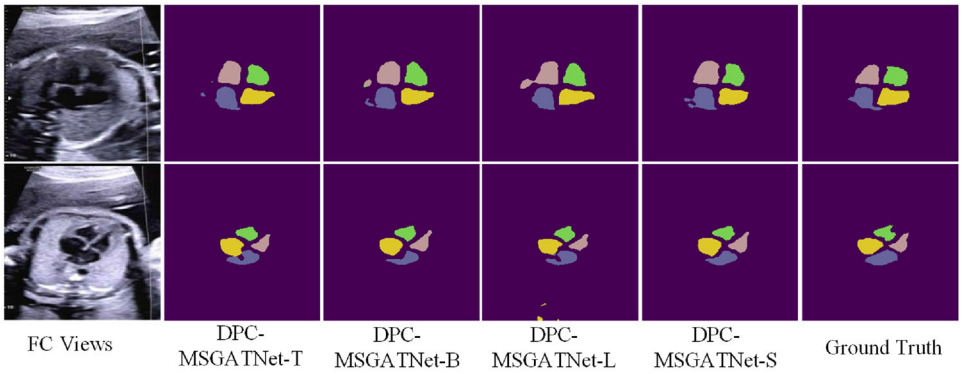

Fig. 9 Visual segmentation comparison of the layers ablation study in our DPC-MSGATNet

Table 6 Inference time comparison of our DPC-MSGATNet against SOTA methods

Methods	Inference Time (s)
U-Net [16]	**0.2494**
Res-UNet [19]	0.2507
U-Net++ [17]	0.2513
Attention U-Net [18]	0.2549
Axial-Attention U-Net [37]	0.4398
Gated-Axial-Attention U-Net [33]	0.4448
MedT [33]	0.9604
Global Branch	**0.4869**
DPC-MSGATNet-T	0.6617
DPC-MSGATNet-S	0.8464
DPC-MSGATNet-B	1.0529
DPC-MSGATNet-L	1.2435

Bold metric means that its corresponding method performs best among other SOTA methods

on other downstream tasks, we choose two public medical datasets, GLAS [38], and MonuSeg [39], to test our model. The 7 SOTA methods are also adopted to compare performance with our DPC-MSGATNet. Table 7 shows the quantitative comparison of our DPC-MSGATNet with the 7 SOTA mentioned above methods. As can be seen, the CNN-based models outperform the one-branch transformer-based attention models, Axial-Attention U-Net [37], and Gated-Axial-Attention U-Net [33], on the GLAS [38], and MonuSeg [39] datasets, which is quite different from their performance on the fetal FC view dataset. The two public datasets have fewer images than the fetal FC view dataset. From this point, one-branch transformer-based attention models are inferior to CNN-based baselines with small-scale training data. Furthermore, with the assistance of the gated axial attention mechanism and multi-scale branches, the MedT [33], and our DPC-MSGATNet perform better than other methods. It is noteworthy that our proposed DPC-MSGATNet outperforms the MedT [33] by a large margin on both GLAS [38] and MonuSeg [39] datasets, which attributes to our extraordinary global branch architecture and IDPCGAT. Figures 10 and 11 show the visual segmentation performance of our DPC-MSGATNet and 7 SOTA methods on GLAS and MonuSeg datasets. We can see that the visual results are consistent with the above description, proving that our model has solid generalized performance on other downstream tasks.

Conclusion

In this work, we propose a DPC-MSGATNet to precisely segment the fetal cardiac four chambers, which can assist

Table 7 Quantitative performance comparison of our DPC-MSGATNet with SOTA methods on two public datasets

Methods	GLAS dataset		MonuSeg dataset	
	F1 (%)	IoU (%)	F1 (%)	IoU (%)
Axial-Attention U-Net [37]	78.00 ± 0.25	65.36 ± 0.38	76.19 ± 1.21	61.86 ± 1.51
Gated-Axial-Attention U-Net [33]	78.90 ± 0.99	65.81 ± 0.46	76.78 ± 0.41	62.61 ± 0.45
U-Net [16]	79.19 ± 1.04	67.42 ± 0.68	79.30 ± 0.53	65.84 ± 0.52
U-Net++ [17]	79.54 ± 0.78	67.45 ± 0.50	79.54 ± 0.59	66.23 ± 0.15
Res-UNet [19]	79.97 ± 0.76	67.86 ± 0.50	79.57 ± 0.63	66.31 ± 0.11
Attention U-Net [18]	81.45 ± 0.61	69.75 ± 0.54	79.76 ± 0.72	66.48 ± 0.99
MedT [33]	82.72 ± 0.41	72.05 ± 0.48	80.65 ± 0.88	67.34 ± 0.81
DPC-MSGATNet	**85.22 ± 0.17**	**75.29 ± 0.26**	**82.61 ± 1.18**	**70.69 ± 1.68**

Bold metric means that its corresponding method performs best among other SOTA methods

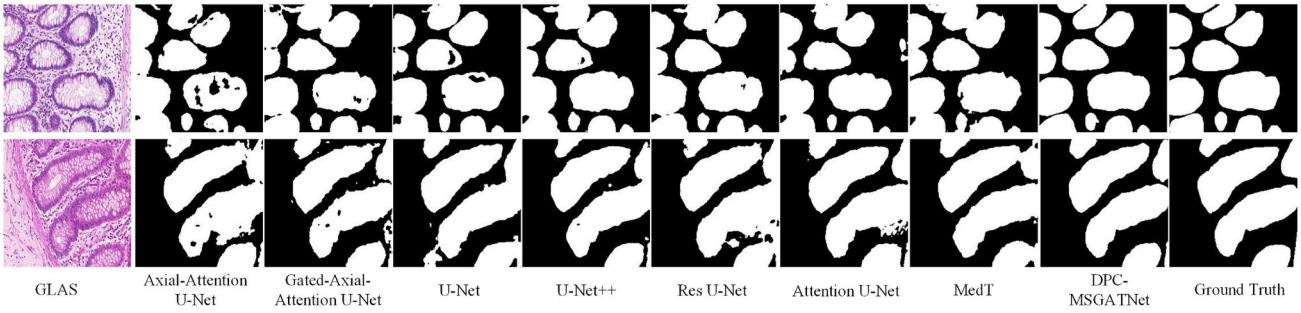

Fig. 10 Visualization segmentation comparison of our DPC-MSGATNet with SOTA methods on GLAS

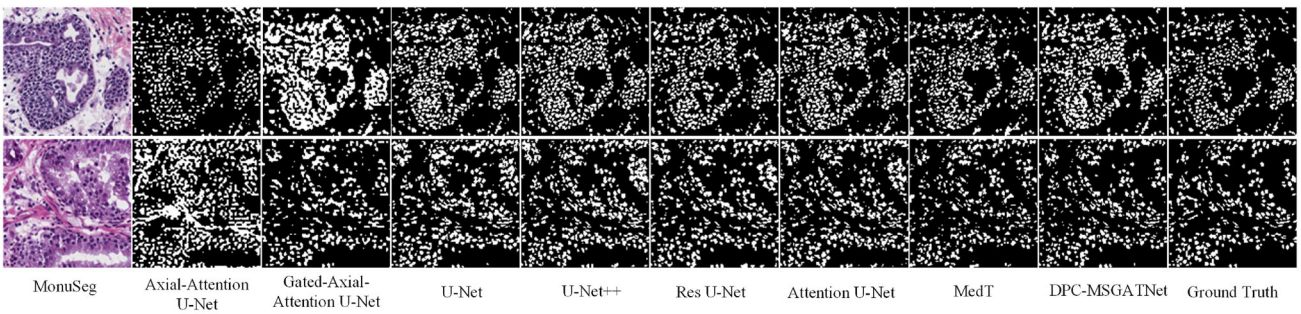

Fig. 11 Visualization segmentation comparison of our DPC-MSGATNet with SOTA methods on MonuSeg

clinicians in analyzing cardiac morphology and promote fetal CHD diagnosis. The DPC-MSGATNet includes a global and a local branch, which operates on the whole image and image patches, and captures global and local visual cues to obtain multi-scale representations from fetal FC views. Moreover, we propose an IDPCGAT to enhance the interactions between global and local branches. The multi-scale representations from the two branches can complement each other, capture the same region's salient features, and suppress feature responses to retain only the activations associated with specific targets. Extensive experiments demonstrate that our DPC-MSGATNet performs better than the seven SOTA CNNs- and transformer-based methods by a large margin in terms of both F1 and IoU scores on the fetal FC views dataset. In addition, we also adopt two public medical datasets (e.g., GLAS and MonuSeg) to verify the generalization of our DPC-MSGATNet, achieving the SOTA segmentation performance.

Our DPC-MSGATNet still has two shortcomings: (1) the model is not a lightweight network, which will affect the model's efficiency in the actual deployment. (2) The model requires labeled data to conduct supervised training. In general, the FC views' annotation is complex and needs experienced cardiologists to spend a long time annotating the dataset.

In the future, we will focus on design principles to reduce the computational cost of the model while maintaining its accuracy. The multilayer perceptron will be combined with convolutional layers to capture effective representations of fetal cardiac contours. Then, the new methods will reduce the number of parameters and speed up the inference time while achieving good performance on segmentation. Furthermore, we will train the model in a semi-supervised strategy, drastically reducing reliance on labeled data.

Acknowledgements The authors are thankful to the Innovation fund project for graduate students of China University of Petroleum (East China) (No. 22CX04036A), and National Natural Science Foundation of China (No. 61873281) for their financial supports.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this work.

Appendix: Abbreviations

As shown in Table 8, we provide the abbreviations of the professional terms used in this work.

Table 8 Abbreviations of their professional terms

Abbreviations	Descriptions
CT	Computed tomography
MRI	Magnetic resonance imaging
US	Ultrasound
FC	Four chamber
LA	Left atrium
LV	Left ventricle
RA	Right atrium
RV	Right ventricle
CNNs	Convolutional neural networks
FCN	Fully convolution network
AsAo	Ascending aorta
PA	Pulmonary artery
MoLV	Myocardium of the LV
TR	Thorax
ED	Epicardium
LVW	Left ventricular wall
RVW	Right ventricular wall
IS	Interatrial septum
LL	Left lung
RL	Right lung
SN	Spine
RB	Ribs
DAO	Descending aorta
IVS	Inter-ventricular septum

References

1. Wang L, Nie H, Wang Q et al (2019) Use of magnetic resonance imaging combined with gene analysis for the diagnosis of fetal congenital heart disease. BMC Med Imaging 19:12
2. Pan S (2019) Exploration and prospect of interventional therapy for fetal congenital heart diseases in china. J Intervent Radiol 28(10):915–920
3. Reddy UM, Filly RA, Copel JA (2008) Prenatal imaging: ultrasonography and magnetic resonance imaging. Obstet Gynecol 112(1):145–157
4. Shi C, Song L, Li Y et al (2020) Value of four-chamber view of the fetal echocardiography for the prenatal diagnosis of congenital heart disease. Chin J Obst Gynecol 37(7):385–387
5. Pan S, Luo G (2020) Application prospect of medical artificial intelligence in fetal echocardiography. Chin J Pract Pediat 35(11):850–853
6. Noble A, Boukerroui D (2006) Ultrasound image segmentation: a survey. IEEE Trans Med Imaging 25(8):987–1010
7. Leclerc S, Smistad E, Pedrosa J et al (2019) Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Trans Med Imaging 38(9):2198–2210
8. Xu L, Liu M, Shen Z et al (2020) Dw-net: a cascaded convolutional neural network for apical four-chamber view segmentation in fetal echocardiography. Comput Med Imaging Graph 80:101690
9. Rahmatullah B, Sarris I, Papageorghiou A et al (2011) Quality control of fetal ultrasound images: detection of abdomen anatomical landmarks using adaboost. In: From Nano to Macro, IEEE International Symposium on Biomedical Imaging, pp 6–9
10. Rahmatullah B, Papageorghiou AT, Noble JA (2012) Integration of local and global features for anatomical object detection in ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp 402–409
11. Maraci MA, Napolitano R, Papageorghiou A, et al (2014) Searching for structures of interest in an ultrasound video sequence. In: International Workshop on Machine Learning in Medical Imaging, pp 133–140
12. Zheng Y, Barbu A, Georgescu B et al (2008) Four-chamber heart modeling and automatic segmentation for 3-d cardiac ct volumes using marginal space learning and steerable features. IEEE Trans Med Imaging 27(11):1668–1681
13. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
14. Xie S, Girshick R, Dollar P, et al (2017) Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 5987–5995
15. Qiao S, Pang S, Luo G, et al (2022) Flds: an intelligent feature learning detection system for visualizing medical images supporting fetal four-chamber views. IEEE J Biomed Health Inform 26(10)
16. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 234–241
17. Zhou Z, Siddiquee MMR, Tajbakhsh N (2020) Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 39(6):1856–1867
18. Oktay O, Schlemper J, Folgoc L, et al (2018) Attention u-net: learning where to look for the pancreas. In: International Conference on Medical Imaging with Deep Learning
19. Zhang Z, Liu Q, Wang Y (2018) Road extraction by deep residual u-net. IEEE Geosci Remote Sens Lett 15(5):749–753
20. Li X, Chen H, Qi X et al (2018) H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE Trans Med Imaging 37(12):2663–2674
21. Huang H, Lin L, Tong R, et al (2020) Unet 3+: a full-scale connected unet for medical image segmentation. In: International Conference on Acoustics, Speech and Signal Processing, 1055–1059
22. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Ann Conf Neural Inform Process Syst 30:5998–6008
23. Shaw P, Uszkoreit J, Vaswani A (2018) Self-attention with relative position representations. In: The 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 2, pp 464–468

24. Nam H, Ha JW, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 299–307
25. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. https://openreview.net/forum?id=YicbFdNTTy
26. Zhu X, Su W, Lu L, et al (2021) Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations https://openreview.net/forum?id=gZ9hCDWe6ke
27. Wang W, Xie E, Li X, et al (2021) Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: IEEE International Conference on Computer Vision, 548–558
28. Chen H, Wang Y, Guo T, et al (2021) Pre-trained image processing transformer. In: IEEE Conference on Computer Vision and Pattern Recognition, 12294–12305
29. Xu Y, Zhang Q, Zhang J et al (2021) Vitae: vision transformer advanced by exploring intrinsic inductive bias. Annu Conf Neural Inform Process Syst 34:28522–28535
30. Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. IEEE Int Conf Comput Vis: 9992–10002
31. Chen J, Lu Y, Yu Q, et al (2021) Transunet: transformers make strong encoders for medical image segmentation. arXiv:2102.04306
32. Zhang Y, Liu H, Hu Q (2021) Transfuse: fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 14–24
33. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM (2021) Medical transformer: gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 36–46
34. Karimi D, Vasylechko S, Gholipour A (2021) Convolution-free medical image segmentation using transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention pp 78–88
35. Zhang Y, Higashita R, Fu H (2021) A multi-branch hybrid transformer network for corneal endothelial cell segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 99–108
36. Wang W, Chen C, Ding M (2021) Transbts: multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp 109–119
37. Wang H, Zhu Y, Green B, et al (2020) Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. Eur Conf Comput Vis: 108–126
38. Sirinukunwattana K, Pluim JP, Chen H et al (2017) Gland segmentation in colon histology images: the glas challenge contest. Med Image Anal 35:489–502
39. Kumar N, Verma R, Anand D et al (2019) A multi-organ nucleus segmentation challenge. IEEE Trans Med Imaging 39(5):1380–1391
40. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Tran Pattern Anal Mach Intell 39(4):640–651
41. Shah S, Ghosh P, Davis L, et al (2018) Stacked u-nets: a no-frills approach to natural image segmentation. arXiv:1804.10343
42. Yu T, Li X, Cai Y, Sun M, Li P (2022) S2-mlp: Spatial-shift mlp architecture for vision. In: IEEE Winter Conference on Applications of Computer Vision, pp 3615–3624
43. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, et al (2021) Mlp-mixer: an all-mlp architecture for vision. Adv Neural Inform Process Syst: 24261–24272
44. Valanarasu JMJ, Patel VM (2022) Unext: Mlp-based rapid medical image segmentation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp 23–33
45. Mo Y, Liu F, Mcilwraith D, et al (2018) The deep poincaré map: a novel approach for left ventricle segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp 561–568
46. Xue W, Li J, Hu Z et al (2021) Left ventricle quantification challenge: a comprehensive comparison and evaluation of segmentation and regression for mid-ventricular short-axis cardiac mr data. IEEE J Biomed Health Inform 25(9):3541–3553
47. Avendi MR, Kheradvar A, Jafarkhani H (2017) Automatic segmentation of the right ventricle from cardiac mri using a learning-based approach. Magn Reson Med 78(6):2439–2448
48. Chen J, Zhang H, Zhang W et al (2018) Correlated regression feature learning for automated right ventricle segmentation. IEEE J Transl Eng Health Med 6:1–10
49. Duan J, Bello G, Schlemper J et al (2019) Automatic 3d bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach. IEEE Tran Med Imaging 38(9):2151–2164
50. Wang C, MacGillivray T, Macnaught G et al (2019) A two-stage u-net model for 3d multi-class segmentation on full-resolution cardiac data. Stati Atlas Comput Models Heart Atrial Segment LV Quantif Challeng 11395:191–199
51. Yu L, Guo Y, Wang Y et al (2017) Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks. IEEE Trans Biomed Eng 64(8):1886–1895
52. Yang T, Han J, Zhu H, et al (2020) Segmentation of five components in four chamber view of fetal echocardiography. Int Symp Biomed Imaging: 1962–1965
53. An S, Zhu H, Wang Y et al (2021) A category attention instance segmentation network for four cardiac chambers segmentation in fetal echocardiography. Comput Med Imaging Graph 93:101983
54. Guo L, Lei B, Chen W et al (2021) Dual attention enhancement feature fusion network for segmentation and quantitative analysis of paediatric echocardiography. Med Image Anal 71:102042
55. Hu Y, Xia B, Mao M et al (2020) Aidan: an attention-guided dual-path network for pediatric echocardiography segmentation. IEEE Access 8:29176–29187
56. Moradi S, Oghli MG, Alizadehasl A et al (2019) Mfp-unet: a novel deep learning based approach for left ventricle segmentation in echocardiography. Phys Med 67:58–69
57. Pu B, Lu Y, Chen J et al (2022) Mobileunet-fpn: a semantic segmentation model for fetal ultrasound four-chamber segmentation in edge computing environments. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2022.3182722
58. Lin TY, Dollar P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. IEEE Conf Comput Vis Pattern Recogn: 2117–2125
59. Howard AG, Zhu M, Chen B, Kalenichenko D, et al (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861
60. Zhao C, Xia B, Guo L, Du J, et al (2021) Multi-scale wavelet network algorithm for pediatric echocardiographic segmentation via feature fusion. IEEE Int Symp Biomed Imaging: 1402–1405